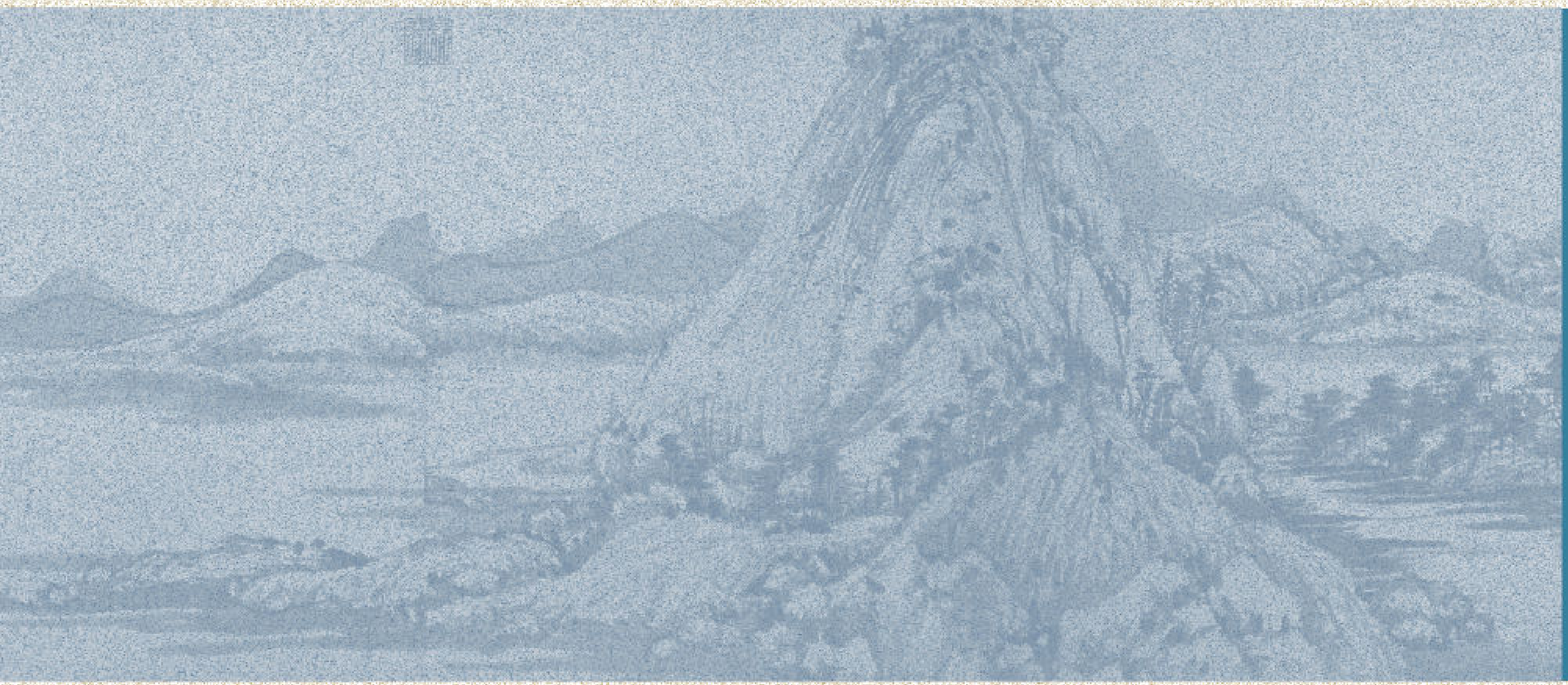
普通高校"十三五"规划教材·统计学系列

应用统计学
——以Excel为分析工具（第2版）



宋廷山
王　坚
刁艳华
郭思亮 ◎ 主编

清华大学出版社

普通高校“十三五”规划教材·统计学系列

应用统计学
——以 Excel 为分析工具（第 2 版）

宋廷山　王　坚　刁艳华　郭思亮　主编

清华大学出版社
北　京

内 容 简 介

本书立足于一级学科的统计学，其内容涵盖了描述统计和推断统计。按照一个完整的统计工作过程的脉络设计了内容：统计学总论、统计设计、统计调查、统计整理和统计分析（平均分析、时间序列分析、指数分析、抽样推断、相关与回归分析和统计分析报告）。

本书为满足普通读者对统计学的基本需要，重点介绍了统计学的基础部分——描述统计（第一章至第七章）；为满足较高层次读者的需要，适当增加了推断统计学的内容（第八章至第十章）。

本书侧重于统计思想的介绍，避开了深奥的数学证明，对于复杂的统计计算通过常用的计算机应用软件 Excel 2013 来实现，以提高读者运用统计方法分析和解决问题的能力。

本书可作为高等院校经济管理类各专业讲授和学习统计学课程的教材或参考书，也可作为政府和企事业单位从事实际统计工作的参考书。

图书在版编目(CIP)数据

应用统计学：以 Excel 为分析工具/宋廷山，王坚，刁艳华，郭思亮主编．—2 版．—北京：清华大学出版社，2018
（普通高校"十三五"规划教材·统计学系列）
ISBN 978-7-302-48212-3

Ⅰ．①应…　Ⅱ．①宋…　②王…　③刁…　④郭…　Ⅲ．①应用统计学－高等学校－教材
Ⅳ．①C819

中国版本图书馆 CIP 数据核字(2017)第 207292 号

责任编辑：张　伟
封面设计：汉风唐韵
责任校对：王荣静
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62770177-4506
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：20　　**字　　数**：468 千字
版　　次：2012 年 5 月第 1 版　　2018 年 1 月第 2 版　　**印　　次**：2018 年 1 月第1 次印刷
印　　数：1～4000
定　　价：42.00 元

产品编号：065555-01

第2版前言

本书自2012年5月第1版与读者见面已经5年有余,连续印刷5次,我们全体作者在感到欣慰的同时,由衷地感谢那些关注、关心、支持我们的读者、出版社的有关人员和老师、领导、同仁,我们感觉身上的担子更重了,责任感和使命感更强烈了。正是你们的关心、鞭策、鼓励,促使我们高标准地要求自己,使本书不断完善、优化和提高。为了更好地满足读者的需求,更优质地服务于读者,让读者从中受益更多,不让我们的读者、出版社的有关人员和老师们、领导们、同仁们失望,我们本着"创优质课程、建优质教材、育优质人才"的宗旨,根据使用者的反馈意见,组织力量对其进行了如下修订:

第一,删除了综合评价的内容,增加了统计分析报告的内容。由于大多数高校不讲授综合评价这部分内容;又由于统计工作过程的最后一个环节是统计分析,而统计分析的最终成果是统计分析报告,如何撰写统计分析报告就需要学生掌握,所以我们本次修订,就将统计综合评价一章改成统计分析报告。

第二,每一章开始都增加了统计知识ABC。增加的统计知识ABC更能体现本章的内容,能起到引导案例的作用,引导和启发读者学习相关章节统计知识的兴趣和积极性。在第三章第四节调查问卷中增加了问卷的预评估内容,使问卷调查更加完整,增强了实用性。

第三,更新了最新数据,第1版中涉及的实际数据都替换了最新的数据。对于Excel实现部分,第1版使用Excel 2007版本,第2版全部替换成较新的Excel 2013版本,都给出了较详细的操作说明和结果解释。

第四,将第1版中的第九章相关与关联分析和第十章回归分析内容进行压缩合并成了一章,使本书体系更加紧凑。

本书由齐鲁师范学院宋廷山、青岛农业大学王坚、山东协和学院刁艳华、齐鲁师范学院郭思亮任主编。参加本次修订的有宋廷山(统计知识ABC和本书体系调整)、王坚(更新数据和错误修改)、刁艳华(Excel 2013操作)、郭思亮(统计分析报告和问卷预评估内容)。愿第2版教材能对您的工作和学习有所帮助。同时也衷心地希望使用本书的老师们、同学们和其他读者给予批评指正,我们将不胜感激。

编　者

2017年3月8日于泉城济南

第1版前言

统计学的英文名字是 statistics，最早源于现代拉丁文 statisticum collegium、意大利文 statista、德文 statistik。统计学是一门很古老的学科，一般认为其学理研究始于古希腊的亚里士多德(Aristoteles)时代，迄今已有两千三百多年的历史。

随着社会、经济和科学技术的发展，统计在现代化国家管理、企业管理和社会生活中的地位，越来越重要了，人们的日常生活都离不开统计。英国统计学家哈斯利特说："统计方法的应用是这样普遍，在我们的生活和习惯中，统计的影响是这样巨大，以致统计的重要性无论怎样强调也不过分。"

国务院学位办 2011 年颁布了新的学科目录，学科的第一层次是门类，依次包括哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学、艺术学共 13 个门类。每个门类之下，再分一级学科；一级学科也还可以再往下细分。第七大门类是理学，其中一共有 14 个一级学科，第一是数学，第十四是统计学。调整后的统计学一级学科将原属应用经济学和数学下与统计相关的学科进行了整合，并在一级学科下设有数理统计、社会经济统计、生物卫生统计、金融统计与风险管理和精算、应用统计 5 个二级学科，可以根据各校情况授予经济学学位或者理学学位。

在这种形势下，亟须编写一本反映时代特征的统计学教材。

我们在广泛吸收近年来出版的优秀统计学教材、著作的优点的基础上，将本书定位为应用统计学(以示与概率论与数理统计——理论统计学区别)。全书共设计了 11 章，系统介绍了应用统计学的基本理论和方法。

本书具有以下特点。

(1) 脉络清楚、体系完整、结构紧凑。立足于一级学科的"大统计学"，篇章结构涵盖了描述统计和推断统计。按照一个完整的统计工作过程的脉络来设计篇章结构：统计学总论(第一章)、统计设计(第二章)、统计调查(第三章)、统计整理(第四章)和统计分析。其中统计分析又细分为平均分析(第五章)、时间序列分析(第六章)、指数分析(第七章)、抽样推断(第八章)、相关及关联分析(第九章)、回归分析(第十章)和统计综合评价(第十一章)。

(2) 强化基础、兼顾提高、重在应用。为满足专科生等普通读者对统计学的基本需要，本书的重点放在了统计学的基础部分——描述统计(第一章至第八章)；为满足本科、研究生等较高层次读者的需要，适当增加了推断统计学的内容(第九章至第十一章)，尤其是增设了难度较大、应用水平较高、实用性较强的统计综合评价(分析)内容，兼顾了普及和提高两个方面。在高等院校经济、管理类核心课程中，统计学是比较难学的一门课程，统计学用到的数学知识较多；应用方面的灵活性较强；计算量大且复杂。为解决这一矛

盾,本书致力于将统计学的应用作为出发点和归宿,侧重于统计思想的介绍,避开了深奥的数学证明,对于复杂的统计计算试图通过常用的计算机应用软件 Excel 来实现,力图培养读者的学习兴趣,在树立读者统计意识或统计思想的基础上,能用 Excel 完成统计计算,正确解读统计结果,以提高读者运用统计方法分析和解决问题的能力。

本书由齐鲁师范学院宋廷山教授、青岛农业大学王坚副教授、山东科技大学姜爱萍副教授担任主编;山东协和学院郭思亮、山东青年政治学院刘莹、山东轻工业学院宋志涛担任副主编;青岛农业大学王宝海教授主审。编写提纲由宋廷山教授提出初稿,经主编、副主编讨论形成。参加编写的有:宋廷山(第一章、第二章、第十一章);王坚(第七章);姜爱萍(第五章);王宝海(第六章);郭思亮(第三章、第四章);刘莹(第九章);齐鲁师范学院曲岩(第十章);宋志涛(第八章)。初稿形成后,由副主编进行了修改,最后由主编进行总纂,由王宝海教授进行审校。

本书编写过程中,我们参考了国内外出版的大量本学科教材和专著,敬列于参考文献中。对编写这些著作的学界前辈、专家和同行们,我们表示崇高的敬意和衷心的感谢!

由于水平所限,加之时间紧迫,教材中可能还存在许多我们还没有发现的问题,衷心希望使用本教材的老师、同学和其他读者批评指正,有问题或建议可发电子邮件至 qlnusts@163.com,对于特殊性的问题我们将给予个别答复,对于具有普遍性的问题,将在再版时进行更正和说明。在此我们一并表示感谢!

编　者

2011 年 12 月于泉城济南

目 录

第一章

统计学总论

【统计知识 ABC】

[A]GDP

GDP(国内生产总值)是指一个国家(或地区)所有常住单位在一定时期内生产活动的最终成果。国内生产总值有三种表现形态,即价值形态、收入形态和产品形态。从价值形态看,它是所有常住单位在一定时期内生产的全部货物和服务价值超过同期中间投入的全部非固定资产货物和服务价值的差额,即所有常住单位的增加值之和;从收入形态看,它是所有常住单位在一定时期内创造并分配给常住单位和非常住单位的初次收入分配之和;从产品形态看,它是所有常住单位在一定时期内最终使用的货物和服务价值与货物和服务净出口价值之和。在实际核算中,国内生产总值有三种计算方法,即生产法、收入法和支出法。三种方法分别从不同的方面反映了国内生产总值及其构成。

支出法国内生产总值指一个国家(或地区)所有常住单位在一定时期内用于最终消费、资本形成总额,以及货物和服务的净出口总额,它反映本期生产的国内生产总值的使用及构成。**最终消费**指常住单位在一定时期内对于货物和服务的全部最终消费支出,也就是常住单位为满足物质、文化和精神生活的需要,从本国经济领土和国外购买的货物和服务的支出;不包括非常住单位在本国经济领土内的消费支出。最终消费分为居民消费和政府消费。**居民消费**指常住住户对货物和服务的全部最终消费支出。居民消费按市场价格计算,即按居民支付的购买者价格计算。购买者价格是购买者取得货物所支付的价格,包括购买者支付的运输和商业费用。居民消费除了直接以货币形式购买货物和服务的消费之外,还包括以其他方式获得的货物和服务的消费支出,即所谓的虚拟消费支出。居民虚拟消费支出包括以下几种类型:单位以实物报酬及实物转移的形式提供给劳动者的货物和服务;住户生产并由本住户消费了的货物和服务,其中的服务仅指住户的自有住房服务;金融机构提供的金融媒介服务;保险公司提供的保险服务。**政府消费**指政府部门为全社会提供公共服务的消费支出和免费或以较低价格向住户提供的货物和服务的净支出。前者等于政府服务的产出价值减去政府单位所获得的经营收入的价值,政府服务的产出价值等于它的经常性业务支出加上固定资产折旧;后者等于政府部门免费或以较低价格向住户提供的货物和服务的市场价值减去向住户收取的价值。**资本形成总额**指常住单位在一定时期内获得的减去处置的固定资产加存货的变动,包括固定资本形成总额和存货增加。**固定资本形成总额**指常住单位购置、转入和自产自用的固定资产,扣除固定资

产的销售和转出后的价值,分有形固定资产形成总额和无形固定资产形成总额。有形固定资产形成总额包括一定时期内完成的建筑工程、安装工程和设备工器具购置(减处置)价值,以及土地改良、新增役、种、奶、毛、娱乐用牲畜和新增经济林木价值。无形固定资产形成总额包括矿藏的勘探、计算机软件、娱乐和文学艺术品原件等获得减处置。**存货增加**指常住单位存货实物量变动的市场价值,即期末价值减期初价值的差额。存货增加可以是正值,也可以是负值;正值表示存货上升,负值表示存货下降。它包括生产单位购进的原材料、燃料和储备物资等存货,以及生产单位生产的产成品、在制品等存货。**货物和服务净出口**指货物和服务出口减货物和服务进口的差额。出口包括常住单位向非常住单位出售或无偿转让的各种货物和服务的价值;进口包括常住单位从非常住单位购买或无偿得到的各种货物和服务的价值。由于服务活动的提供与使用同时发生,因此服务的进出口业务并不发生出入境现象,一般把常住单位从国外得到的服务作为进口,非常住单位从本国得到的服务作为出口。货物的出口和进口都按离岸价格计算。

收入法国内生产总值包括劳动者报酬、生产税净额、固定资产折旧、营业盈余。**劳动者报酬**指劳动者因从事生产活动所获得的全部报酬。包括劳动者获得的各种形式的工资、奖金和津贴,既包括货币形式的,也包括实物形式的,还包括劳动者所享受的公费医疗和医药卫生费、上下班交通补贴和单位支付的社会保险费等。对于个体经济来说,其所有者所获得的劳动报酬和经营利润不易区分,这两部分统一作为劳动者报酬处理。**生产税净额**指生产税减生产补贴后的余额。生产税指政府对生产单位生产、销售和从事经营活动以及因从事生产活动使用某些生产要素(如固定资产、土地、劳动力)所征收的各种税、附加费和规费。生产补贴与生产税相反,指政府对生产单位的单方面收入转移,因此视为负生产税,包括政策亏损补贴、粮食系统价格补贴、外贸企业出口退税收入等。**固定资产折旧**指一定时期内为弥补固定资产损耗按照核定的固定资产折旧率提取的固定资产折旧,或按国民经济核算统一规定的折旧率虚拟计算的固定资产折旧。它反映了固定资产在当期生产中的转移价值。各类企业和企业化管理的事业单位的固定资产折旧是指实际计提并计入成本费中的折旧费;不计提折旧的政府机关、非企业化管理的事业单位和居民住房的固定资产折旧是按照统一规定的折旧率和固定资产原值计算的虚拟折旧。原则上,固定资产折旧应按固定资产的重置价值计算,但是目前我国尚不具备对全社会固定资产进行重估价的基础,所以暂时只能采用上述办法。**营业盈余**指常住单位创造的增加值扣除劳动者报酬、生产税净额和固定资产折旧后的余额。它相当于企业的营业利润加上生产补贴,但要扣除从利润中开支的工资和福利等。

生产法国内生产总值是指各行业增加值之和。其计算公式为

$$\text{生产法国内生产总值} = \sum(\text{总产出} - \text{中间投入}) = \sum \text{各行业增加值}$$

(资料来源:http://www.stats.gov.cn/tjzd/tjzbjs/t20020327_14293.html.)

[B]GNP 或 GNI

国民生产总值(GNP)又称国民总收入(gross national income,GNI),指一个国家(或地区)所有常住单位在一定时期内收入初次分配的最终结果。一国常住单位从事生产活动所创造的增加值在初次分配中主要分配给该国的常住单位,但也有一部分以生产税及

进口税(扣除生产和进口补贴)、劳动者报酬和财产收入等形式分配给非常住单位;同时,国外生产所创造的增加值也有一部分以生产税及进口税(扣除生产和进口补贴)、劳动者报酬和财产收入等形式分配给该国的常住单位,从而产生了国民生产总值的概念。它等于国内生产总值加上来自国外的净要素收入。与国内生产总值不同,国民生产总值是个收入概念,而国内生产总值是个生产概念。两者之间的数量关系为

国民生产总值＝国内生产总值＋来自国外的要素收入净额
＝国内生产总值＋来自国外的要素收入
－支付给国外的要素收入

(资料来源:http://www.stats.gov.cn/tjzd/tjzbjs/t20020327_14293.html.)

[C]社会消费品零售总额

社会消费品零售总额指国民经济各行业直接售给城乡居民和社会集团的消费品总额。它是反映各行业通过多种商品流通渠道向居民和社会集团供应的生活消费品总量,是研究国内零售市场变动情况、反映经济景气程度的重要指标。

社会消费品零售总额包括:①售给城乡居民作为生活用的商品和修建房屋用的建筑材料;②售给社会集团的各种办公用品和公用消费品;③售给机关、团体、学校、部队、企业、事业单位的职工食堂和旅店(招待所)附设专门供本店旅客食用,不对外营业的食堂的各种食品、燃料;企业、单位和国有农场直接售给本单位职工和职工食堂的自己生产的产品;④售给部队干部、战士生活用的粮食、副食品、衣着品、日用品、燃料;⑤售给来华的外国人、华侨、港澳台同胞的消费品;⑥居民自费购买的中、西药品、中药材及医疗用品;⑦报社、出版社直接售给居民和社会集团的报纸、图书、杂志,集邮公司出售的新、旧纪念邮票、特种邮票、首日封、集邮册、集邮工具等;⑧旧货寄售商店自购、自销部分的商品;⑨煤气公司、液化石油气站售给居民和社会集团的煤气灶具和罐装液化石油气;⑩农民售给非农业居民和社会集团的商品。不包括售给国民经济各部门企业、事业单位(包括国有经济的农场)生产经营用的各种原材料、燃料、设备、工具等和售给批发零售贸易业、餐饮业作为转卖用的商品,旧货寄售商店受托寄售卖出的商品,服务业的营业收入,邮局出售邮票的收入,自来水、电力、煤气生产(供应)单位的产品供应收入,也不包括农民之间的商品销售。

(资料来源:http://www.stats.gov.cn/tjzd/tjzbjs/t20020327_14281.html.)

第一节　统计的含义

一、统计的含义

"统计"一词在各种实践活动和科学研究领域中经常出现。然而,不同的人或在不同的场合,对其理解是有差异的。比较公认的看法是,统计有三种含义,即统计活动、统计资料和统计学。

(一)统计活动

统计活动又称统计工作,是指收集、整理和分析统计数据,并探索数据的内在数量规

律性的活动过程。

(二)统计资料

统计资料或称统计数据,即统计活动过程所获得的各种数字资料和其他资料的总称。表现为各种反映社会经济现象数量特征的原始记录、统计台账、统计表、统计图、统计分析报告、政府统计公报、统计年鉴等各种数字和文字资料。

(三)统计学

统计学是指阐述统计工作基本理论和基本方法的科学,是对统计工作实践的理论概括和经验总结。它以现象总体的数量方面为研究对象,阐明统计设计、统计调查、统计整理和统计分析的理论与方法,是一门方法论科学。

《不列颠百科全书》的定义:统计学是收集、分析、表述和解释数据的科学。

《中国大百科全书·数学卷》的定义:统计学是一门科学,它研究怎样以有效的方式收集、整理、分析带随机性的数据,并在此基础上对所研究的问题做出统计性推断,直至对可做出的决策提供依据或建议。

统计学的英文是“statistics”。以单数形式出现时,表示一门科学即统计学;以复数形式出现时,表示统计数据或统计资料。

统计工作、统计资料和统计学之间有着密切的联系。统计工作同统计资料是过程同成果的关系,统计资料是统计工作的直接成果。就统计工作和统计学的关系来说,统计工作属于实践的范畴,统计学属于理论的范畴,统计学是统计工作实践的理论概括和科学总结,它来源于统计实践,又高于统计实践,反过来又指导统计实践,统计工作的现代化同统计科学研究的支持是分不开的。

统计工作、统计资料和统计学相互依存、相互联系,共同构成了一个完整的整体,这就是我们所说的统计。

二、统计的职能

统计是适应国家管理的客观需要而逐步产生和发展起来的。现代化国家管理系统,包括决策系统、执行系统、信息系统、咨询系统、监督系统五个组成部分。国家统计兼有信息、咨询、监督三种系统的职能。

(一)信息职能

信息职能是指国家统计部门根据科学的统计指标体系和统计调查方法,灵敏、系统地采集、处理、传递、存储和提供大量的以数量描述为基本特征的社会经济信息。因而,要不断拓展统计信息的内容,保证统计信息的可靠性,完善统计信息的自动化建设,实现统计信息生产和使用的社会化程度。

(二)咨询职能

咨询职能是指利用已经掌握的丰富的统计信息资料,运用科学的分析方法和先进的

技术手段，深入开展综合分析和专题研究，为科学决策和管理提供可供选择的咨询建议与对策方案。

（三）监督职能

监督职能是指根据统计调查和分析，及时、准确地从总体上反映经济、社会和科技运行的实际状况，并对其实行全面、系统的定量检查、监测和预警，以促进国民经济持续、稳定、协调地发展。

上述三种职能是相互联系、相辅相成的。首先，采集和提供信息是统计最基本的职能，统计的信息职能是保证统计咨询和监督职能的基础与前提；其次，统计咨询职能是统计信息职能的延续和深化，它使采集的信息得以在科学决策、经营管理以及社会实践中发挥作用；最后，统计监督职能则是对信息和咨询职能的进一步拓展，统计监督职能的强化，又必然对信息与咨询职能提出更高的要求，从而促进统计信息与咨询职能的优化。总之，统计的信息、咨询、监督职能彼此依存、相互联系，共同构成了一个完整的有机整体。在发挥各自职能的基础上，对三种职能进行优化和整合，形成合力，就能充分发挥统计在国家现代化管理过程中的作用。

三、统计的任务

统计的基本任务是对国民经济和社会发展情况进行统计调查、统计分析，提供统计资料和统计咨询意见，实行统计监督。国家机关、社会团体、企业事业组织和个体工商户等，必须如实提供统计资料，不得虚报、瞒报、拒报、迟报，不得伪造、篡改。基层群众性自治组织和公民有义务如实提供国家统计调查所需要的情况。

第二节　统计学的研究对象和研究方法

一、统计学的研究对象

通过统计学的含义可以看出，统计学的研究对象是现象总体的数量方面，包括数量表现、数量关系和数量界限。

统计学的研究对象具有以下几个特点。

（1）总体性，统计是研究总体的，即群体现象。

（2）数量性，事物能用数量表现的方面和能用数量表现的事物都可以进行统计研究。

（3）差异性，统计所研究的总体内部是有差异的。

（4）具体性，统计研究的不是抽象的量，而是有一定质的规定性的量。

二、统计学的研究方法

统计学的研究方法主要有大量观察法、试验设计法、统计描述法、统计推断法、统计模型法。

（1）大量观察法。大量观察法是对所研究现象总体的全部或足够多的单位进行调查

并加以综合研究的方法。

(2) 试验设计法。试验设计法是通过设计试验取得所要研究的数据的方法。试验设计要遵循的原则:①重复性原则,在相同条件下重复多次试验;②随机化原则,在试验中对试验对象的分配和试验次序是随机安排的;③区组化原则,组内差异大,组间差异小。

(3) 统计描述法。统计描述法是用综合指标、统计表、统计图等形式描述研究总体现象的数量特征的方法。包括统计分组法和综合指标法。

(4) 统计推断法。统计推断法指在一定的置信标准要求下,由样本信息推断总体数量特征的归纳推理方法。包括参数估计、假设检验、方差分析等方法。

(5) 统计模型法。统计模型法是根据统计资料,运用统计方法,对研究现象的结构或过程建立一种统计表达式,进行有关分析的方法。统计模型一般包括四个基本因素:变量、关系式、模型参数和随机项。

第三节　统计研究过程

和人类其他所有认识活动一样,统计研究也要经过一个由现象到本质、由矛盾的特殊性到矛盾的普遍性、由感性认识到理性认识的不断深化的过程。从具体的统计认识活动来看,统计工作是由统计设计、统计调查、统计整理、统计分析等环节组成的。

一、统计设计

统计设计是指根据统计研究对象的性质和研究目的,对统计工作的各个方面和各个环节所做的全面部署和安排。统计设计的最终结果表现为各种标准、规定、制度、方案和办法,如统计分类标准、目录、统计指标体系、统计报表制度、统计调查方案、普查办法、统计整理或汇总方案等。

统计工作是一项高度集中统一和科学性很强的工作,无论是统计总体范围、统计指标的口径和计算方法,还是统计分类和分组的标准,都必须统一,不能各行其是。只有科学地进行统计设计,才能做到统一认识、统一步骤、统一行动,使整个统计工作有秩序地、协调地进行,从而从根本上保证统计工作和统计资料的质量。因此,统计设计是统计工作的先导。

统计设计的主要内容有:统计指标和指标体系的设计、统计分类和统计分组的设计、统计表的设计、统计资料收集方法的设计、统计工作各个部门和各个阶段的协调与联系、统计力量的组织与安排等。

二、统计调查

统计调查是根据统计调查方案的要求,采用各种调查组织形式和调查方法,有组织、有计划地对所研究总体的各个单位进行观察、登记,准确、及时、系统、完整地收集统计资料的过程。

统计调查是统计认识活动由定性认识过渡到定量认识的阶段,这个阶段所收集的资料是否客观、周密、系统、及时,直接影响到统计整理的好坏,关系到统计分析结论的正确

性，决定着整个统计工作的质量。所以，统计调查是整个统计工作的基础。

三、统计整理

统计整理是根据统计研究的目的和任务，对统计调查阶段所取得的原始资料进行审核、分组和汇总，将分散的、零星的、反映总体单位特征的资料转化为反映各组和总体数量特征的综合资料的过程。

统计整理是将对总体单位特征的认识过渡到对总体数量特征的认识的桥梁和纽带，它既是统计调查的继续，又是统计分析的必要前提，在统计工作中，处于中间环节，起着承上启下的作用，是沟通统计调查和统计分析的纽带与桥梁。

四、统计分析

统计分析是指在统计调查和统计整理的基础上，用科学的分析方法，对所研究的现象总体进行全面、系统的数量分析，认识和揭示事物的本质与规律性，进而向有关单位和部门提出咨询建议的统计工作过程。统计分析是统计工作的最后阶段，也是统计发挥信息、咨询和监督职能的关键阶段。

从认识论的角度来说，统计设计属于对社会经济现象进行的定性认识；统计调查和统计整理，是实现对事物个体特征过渡到对总体数量特征认识的关键环节，属于定量认识的范畴；统计分析则是运用统计方法对资料进行比较、判断、推理和评价，揭示社会经济现象的本质和规律性的重要阶段。统计设计、统计调查、统计整理和统计分析的有机统一，体现了统计要在质与量的辩证统一中研究社会经济现象总体数量特征的原则要求。

第四节　统计学的基本概念

在论述统计学的理论与方法的过程中，要运用一些专门的概念，熟悉这些概念是掌握统计学的基础。

一、总体与样本

（一）总体

凡是客观存在、在某一共同性质基础上结合起来的许多个别事物的整体，叫作统计总体（简称总体）。例如，要研究某地区非公有制工业企业的生产经营情况，那么该地区全部非公有制工业企业就构成了一个总体，统计设计、统计调查、统计整理和统计分析则都要围绕这一对象来进行；再如，要研究我国的人口状况，则全国人口就构成了一个统计总体，从设计普查方案、普查登记、资料汇总到最后公布普查数据等，也都要围绕这一对象来进行。

需要注意的问题是，在统计研究过程中，统计研究的目的和任务居于支配与主导地位，是我们考虑一切问题的出发点。一方面，统计总体取决于统计研究的目的和任务，有

什么样的研究目的就要求有什么样的统计总体与之相适应;另一方面,统计研究方法、步骤等也要体现统计研究的目的要求。

统计总体分为有限总体和无限总体两种类型。有限总体是指总体中的总体单位数可以计数或穷尽的总体。例如,一个企业的全体职工、一个国家的全部人口等都是有限总体。如果总体中的单位数是无穷大的,或准确地度量它的单位数是不经济或没有必要的,这样的总体称为无限总体。例如,在连续生产的生产线上产出的全部零件数,一片树林中生长的林木数,江河湖海中生长的鱼的尾数,等等。

划分有限总体和无限总体对于统计工作的意义在于可以帮助我们进行统计调查方法设计。在统计调查方法体系中,全面调查是对调查对象(总体)的全部单位无一遗漏地进行调查,而非全面调查则是对调查对象中的一部分单位进行调查,然后再估计总体指标的方法。很显然,对于有限总体,我们既可以使用全面调查,也可以使用非全面调查,但对于无限总体就只能使用非全面调查。

统计总体具有同质性、大量性和变异性三个特点。

(1) 同质性。所谓同质性是指构成统计总体的各个单位必须在某些方面而且至少在一个方面具备某种共同的性质。同质性是构成统计总体的前提。

(2) 大量性。大量性是指统计总体是由总体的全部单位组成,只有一个单位的统计总体是不存在的。当然,研究目的不同,统计总体就不同,总体中所包含的总体单位的数量也就不同,一个统计总体究竟包含多少总体单位,最终取决于统计研究的目的。

统计总体具有大量性的特点,这是由统计研究对象决定的。如前所述,社会经济统计学的研究对象是大量社会经济现象总体的数量特征,因此,我们只能而且必须将多个具有某种共同性质的单位组合成一个完整的整体,作为统计研究的具体对象。

(3) 变异性。简言之,变异就是事物之间的差别或不同。从统计研究的角度来说,变异性是指构成统计总体的各个单位之间存在的差别。

在此,有三个问题需要特别说明:首先,变异是客观的,没有变异的事物是不存在的;其次,变异对于统计非常重要,没有变异就没有统计,这是因为,如果总体单位之间不存在变异,我们只需要了解一个总体单位的资料就可以推断总体情况了;最后,变异性和同质性之间相互联系、相互补充,是辩证统一的关系。用同质性否定变异性或用变异性否定同质性都是错误的。

(二) 样本

统计研究最终是要确定总体的数量特征,但是有时总体的单位数很多,甚至无限,不可能或无必要对每个总体单位都做调查。这时,就要借助样本来研究总体了。所谓样本,就是按照一定的概率从总体中抽取并作为总体代表的一部分总体单位的集合体。也有学者称总体为母体,样本为子样。但是,绝对不允许将统计总体叫作“全及总体”,样本叫作“样本总体”,这类叫法十分不规范。

样本是统计学中非常重要的概念,对这一概念的理解要注意三方面问题:其一,构成某一样本的每一单位都必须取自某一特定的统计总体,不允许该总体之外的单位介入该

总体的样本。其二，样本单位的抽取应是按一定的概率进行的，而具体样本的产生应是随机的，因此必须排除人为的主观因素对样本单位抽取和样本生成的干扰。其三，样本是母体的代表，带有母体的信息，因而能够推断母体；然而，样本只是母体的一个子集，且具有随机性，故由样本去推断总体会产生代表性误差。其实，如何从母体中抽取子样，怎样控制样本对总体的代表性误差，是推断统计学研究的主要问题。

二、总体单位与标志

（一）总体单位

构成统计总体的个别单位称为统计总体单位，简称总体单位。例如，全部工业企业中的每个企业、全国人口中的每个人也是总体单位。

总体和总体单位的关系是整体同个体、集合同元素的关系，两者相互依存、相互联系，不存在没有总体的总体单位，也不存在没有总体单位的统计总体。

总体和总体单位的具体形式随着统计研究目的的不同而不同，可以是人，也可以是物，还可以是组织（企业或家庭）或时间、空间、行为等。

总体和总体单位的关系不是一成不变的，随着研究目的的变动，两者可以相互转化。在一定研究目的下，一个事物可以作为总体而存在，然而当研究目的发生变化后，这个事物可能就成为总体单位了。例如，研究我国电子工业的发展情况，那么电子工业行业的所有企业就是一个统计总体，每个企业就是一个总体单位；而要研究一个企业的生产经营情况，那么一个企业就构成了统计总体；如果要研究整个国民经济的发展情况，国民经济所有行业组成统计总体，而其中的电子工业行业又变成了总体单位。

（二）标志

统计是从对个体的观察开始，逐步过渡到对总体数量特征的认识的。标志是指说明总体单位特征的名称，如一个企业作为总体单位，这个企业的“所有制类型”“生产能力”“年产量”“销售收入”“职工人数”“工资总额”等都是标志。

标志分为品质标志和数量标志两种类型。品质标志是说明总体单位属性特征的名称，如工人的“性别”“民族”“工种”等，品质标志只能用文字而不能用数值表示。数量标志是说明总体单位数量特征的名称，如工人的“工资额”“工龄”“年龄”等，数量标志是用数值表示的。

统计中反映总体单位特征的标志很多，如果按总体单位在标志上的具体表现是否存在差异来看，可分为不变标志和可变标志。当各个总体单位在某一标志上的具体表现都相同时，则为不变标志。不变标志体现总体的同质性。组成一个总体的各个总体单位必须有一个或几个不变标志，不变标志是使许多个别单位组合成为总体的前提。例如，以全国国有大中型机械工业企业为总体，这里的各工业企业均有所有制、企业规模、工业部门这三个不变标志。一个企业如果不具备这三者任何一个方面特征的话，就不能成为这个总体的一个单位。可变标志是指具体表现在总体各个单位上不相同或不完全相同的那些标志。一般来说，组成总体的各个总体单位具有许多可变标志。例如，把全部国有大型机

械工业企业作为一个统计总体,那么厂址、隶属关系、职工人数、资金额、生产能力、工业增加值、工业总产值、劳动生产率、平均工资、利税额等就是可变标志。和标志相联系的另外一个概念是标志表现。所谓标志表现,是指总体单位特征在某一标志上的具体体现。如一个人的性别是“男”、年龄“50 岁”、民族“汉族”;某企业是“股份制公司”、年产值“10 亿元”。同标志一样,标志表现也分为品质标志表现和数量标志表现两种类型。

在引入标志表现这一概念之后,变异也可以表述为:统计总体单位之间在某一标志上具有不同标志表现的现象。

三、指标与指标体系

(一)指标

指标(统计指标)是说明现象总体数量特征的概念或范畴。例如,要表明某地区全部工业企业这个总体的数量特征,其数量表现可以有:该地区 2016 年年底工业企业单位数 1.2 万个,全年工业总产值 100 亿元,职工人数 200 万人,人均收入 50 000 元,总产值比上年增长 10%等,这些都是统计指标。

标志和指标是两个既有区别又有联系的统计学概念。

两者的主要区别是:①标志是说明总体单位特征的,而指标是说明总体特征的;②标志中的数量标志可以用数值表示,而品质标志不能用数值表示。所有的统计指标都是用数值表示的,不存在不能用数值表示的统计指标。

两者的联系表现在:①统计指标的数值是由标志表现汇总得到的;②在一定条件下(研究目的的调整),指标和标志之间可以变换。当研究目的发生变化以后,原来的总体转化为总体单位,统计指标也就自然地变为数量标志了,反之亦然。

(二)指标体系

统计指标体系是指由若干相互联系的统计指标构成的有机整体。一个统计指标仅仅从一个侧面反映了社会经济现象,而社会经济现象是多方面的、复杂的,要全面、系统地反映社会经济现象,就必须建立统计指标体系。

四、参数与统计量

(一)参数

用来描述总体特征的概括性数字度量,称为参数。如总体平均数(μ)、总体标准差(σ)、总体比例(π)。总体数据通常是不知道的,所以参数是一个未知的常数。正因为如此,我们才进行抽样,根据样本统计量去估计总体参数。

(二)统计量

用来描述样本特征的概括性数字度量,称为统计量。如样本平均数($\overline{X}$)、样本标准差(s)、样本比例(p)。一旦有了具体的样本,统计量就是已知的了。我们就可以根据样本统计量去估计总体参数了。如用样本平均数($\overline{X}$)去估计总体平均数(μ);用样本标准差(s)

去估计总体标准差(σ);用样本比例(p)去估计总体比例(π)等。

五、统计数据

(一) 数据的计量尺度

在计量学的一般分类方法中,依据对事物计量的精确程度,可将所采用的计量尺度由低级到高级、由粗略到精确分为四个层次,即名类尺度、顺序尺度、区间尺度和比尺度。

1. 名类尺度

名类尺度(亦称分类尺度、列名尺度等)是这样一种品质标志,按照它可对研究客体进行平行的分类或分组,使同类同质、异类异质。例如,按照性别将人口分为男、女两类;按照经济性质将企业分为国有、集体、私营、混合制等。这里的"性别"和"经济性质"就是两种名类尺度。名类尺度是最粗略、计量层次最低的计量尺度,利用它只可测度事物之间的类别差,而不能了解各类之间的其他差别。名类尺度计量的结果表现为某种类别,但为了便于统计处理,如为了计算和识别,也可用不同数字或编码表示不同类别。比如用 1 表示男,0 表示女;用 1 表示国有企业,2 表示集体企业,3 表示私营企业等。这些数字只是不同类别的代码,决不意味着它区分了大小,更不能进行任何数学运算。名类尺度能对事物做最基本的测度,是其他计量尺度的基础。

2. 顺序尺度

顺序尺度(亦称序数尺度、顺位尺度等)是这样一种品质标志,利用它不仅能将事物分成不同的类别,还可确定这些类别的等级差别或序列差别。例如,"产品等级"就是一种测度产品质量好坏的顺序尺度,它可将产品分为一等品、二等品、三等品、次品等;"考试成绩"也是一种顺序尺度,它可将成绩分为优、良、中、及格、不及格等;"对某一事物的态度"作为一种顺序尺度,可将人们的态度分为非常同意、同意、保持中立、不同意、非常不同意等。显然,顺序尺度对事物的计量要比名类尺度精确,但它至多测度了类别之间的顺序,而未测量出类别之间的准确差值。因此,顺序尺度的计量结果只能比较大小,不能进行加、减、乘、除等数学运算。

3. 区间尺度

区间尺度(亦称间隔尺度、等距尺度、定距尺度等)是能测度事物类别或次序之间间距的数量标志,更具体地说,区间尺度可将事物区分为不同类别,对这些类别进行排序,并较准确地度量类别之间数量差距的一种计量尺度。该尺度通常使用自然或物理单位作为度量单位,如收入用人民币"元"度量,考试成绩用"百分制"度量,温度用摄氏或华氏的"度"来度量,重量用"克"度量,长度用"米"度量等。区间尺度的计量结果表现为数值。区间尺度的数值可做加、减法运算,如考试成绩 80 分与 90 分之间相差 10 分,一个地区的温度 20℃与另一个地区的 25℃相差 5℃等。但不能做乘、除法运算。区间尺度没有绝对零点。

4. 比尺度

比尺度(亦称比率尺度)的计量结果也表示为数值,跟区间尺度属同一层次,有时对两

者可不做区分。比尺度这种数量标志不仅能测度各类别的大小和多少,还有一个绝对零点作为起点。这个绝对零点是它跟区间尺度的明显差别,就是说,区间尺度中没有绝对零点,即使其计量值为“0”,这个“0”也是有客观内容的数值,即“0”水平,而不表示“没有”或“不存在”。例如,某个学生统计学的考试成绩为“0”分,这个“0”分是他的统计学的客观成绩,并不表示他没有考试成绩或没有任何统计学知识;一个地区的温度为 0℃,这表示一种温度的水平,并不是说没有温度。而比尺度中绝对零点的“0”,表示“没有”或“不存在”。例如,一个人的身高为“0”米,表示这个人不存在;一个人的收入为“0”,表示这个人没有收入;一个产品的产量为“0”,表示没有这种产品等。现实中,大多数场合人们使用的都是比尺度。

比尺度与上述三种计量尺度相比还有一个特性,就是可以计算数值之间的比值。例如,一个人的月工资收入为 600 元,另一个人的为 300 元,可以得出一个人的收入是另一个人的两倍。但区间尺度由于不存在绝对零点,就只能比较数值差,而不能计算比值。比如,可以说 30℃与 15℃之差为 15℃,而不能说 30℃比 15℃热 1 倍。可见,比尺度可以做加、减、乘、除法运算。

上述四种计量尺度对事物的计量层次是由低级到高级、由粗略到精确,逐步递进的。高层次的计量尺度可以计量低层次计量尺度能够计量的事物,但不能反过来。显然,可以很容易地将高层次计量尺度的计量结果转化为低层次计量尺度的计量结果;将考试成绩的百分制转化为五等级分制就是一例。

(二)数据的类型

1. 按计量尺度分

按照所采用的计量尺度不同,可以将统计数据分为分类数据、顺序数据、区间数据和比数据。统计数据是采用某些计量尺度对事物进行计量的结果,但采用不同的计量尺度会得到不同类型的统计数据。就上述四种计量尺度计量的结果来看,我们可以大体上将统计数据分为两种类型:定性数据和定量数据。定性数据(亦称品质数据)是说明事物的品质特征表现的具体类别,不能用数值表示;因这类数据由名类尺度和顺序尺度计量形成,故又可细分为分类数据和顺序数据。定量数据(亦称数量数据或数值型数据)是说明现象数量特征表现的,能够甚至必须用数值来表现;因这类数据由区间尺度和比尺度计量形成,故又可细分为区间数据和比数据。对不同类型的数据,可采用不同的统计方法来处理和分析,比如,对定性数据一般只采用分组法计算,分析各组的频数或频率,而对定量数据则可用更多的统计方法去处理,计算、分析更多的统计指标或统计量。

2. 按数据的收集方法分

按数据的收集方法分类,可将统计数据分为观测数据和实验数据。观测数据是通过调查或观测而收集到的数据,这类数据是在没有对事物人为控制的条件下而得到的,社会经济现象的统计数据几乎都是观测数据。实验数据是在实验中控制实验对象而收集到的数据,如医药研究试验数据、动植物杂交品种试验数据等。自然科学领域的大多数据都是

试验数据。

3. 按数据与时间关系分

按照数据与时间的关系，可以将统计数据分为截面数据和时间数据。截面数据是指同一时间不同空间上的数据。时间数据是指同一空间不同时间上的数据。将两者结合就形成了混合数据。

六、变量

在统计中，一般把表示现象某种特征的概念称为变量。统计标志和指标都可以称为变量。变量的具体表现称为变量值；统计标志的标志表现和指标数值都是变量值。

变量按其属性不同，分为品质变量和数量变量。如果一个变量的变量值由品质数据来记录，该变量就是品质变量，如“性别”就是个品质变量，其变量值是“男”或“女”；“产品等级”也是个品质变量，它可以表现为“一等品”“二等品”“三等品”“次品”等。如果一个变量的数值由数量数据来记录，该变量就是数量变量或称数字变量，如“产品产量”“商品销售额”“零件尺寸”“年龄”“时间”等都是数量变量，它们可以表现为不同的数值。

数量变量按其变量值是否连续可分为离散变量和连续变量。离散变量的取值是有限的，所有取值都以整数断开，且可一一列举，如“企业数”“产品数量”等就是离散变量。而连续变量的取值是无穷的、连续不断的，不能一一列举，如“年龄”“温度”“零件尺寸”等都是连续变量。

变量按其性质不同分为确定性变量和随机变量。确定性变量的取值唯一确定；随机变量的取值由于受众多、细小、不可控制的偶然因素影响，不唯一确定。

第五节　统计应用软件简介

统计应用软件很多，如 Excel、SPSS（Statistical Package for Social Sciences）、SAS（Statistical Analysis System）、Eviews（Econometrics iews）、Stata、Minitab、S-Plus、Statistica、R 语言、马克威软件等，其功能各有千秋。其中 Excel 应用最为广泛。

Excel 这一单词汉语的意思是优秀，胜过他人。Excel 软件是美国微软公司开发的 Windows 环境下运行的电子表格系统。Excel 具有制表功能、数据计算功能、统计图表功能、统计分析功能。微软公司先后推出了 Excel 1997、Excel 2000、Excel 2002、Excel 2003、Excel 2007、Excel 2010、Excel 2013 等不同版本，随着版本的不断提高，Excel 的数据处理功能和操作的简易性不断加强。

本书以 Excel 2013 版本介绍其统计应用。

由于 Excel 集数据的编辑整理、统计分析、图表绘制于一身，您的计算机只要安装了 Office 软件，就有了 Excel，就能使用 Excel 了；目前的大学生都学习了计算机应用基础，已经有了 Excel 的操作基础；加之 Excel 的统计功能能够满足现有统计学的学习要求，故本书选择了易获得、普及率较高、操作简单的 Excel 应用软件。

本书主要用到 Excel 的数据分析功能（图 1-1）、函数计算功能（图 1-2）和统计制图功能（图 1-3）。

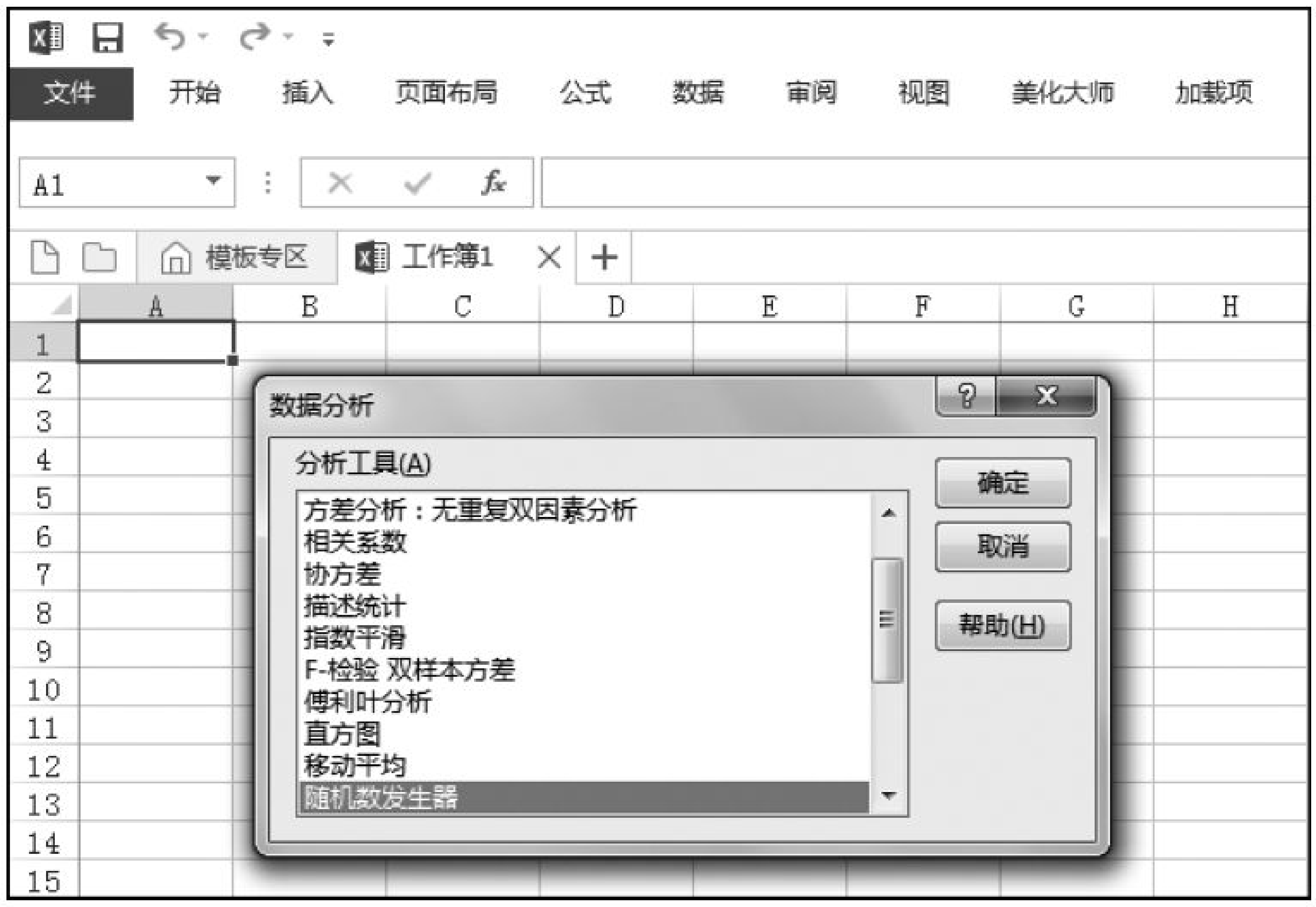


图 1-1 Excel 数据分析功能

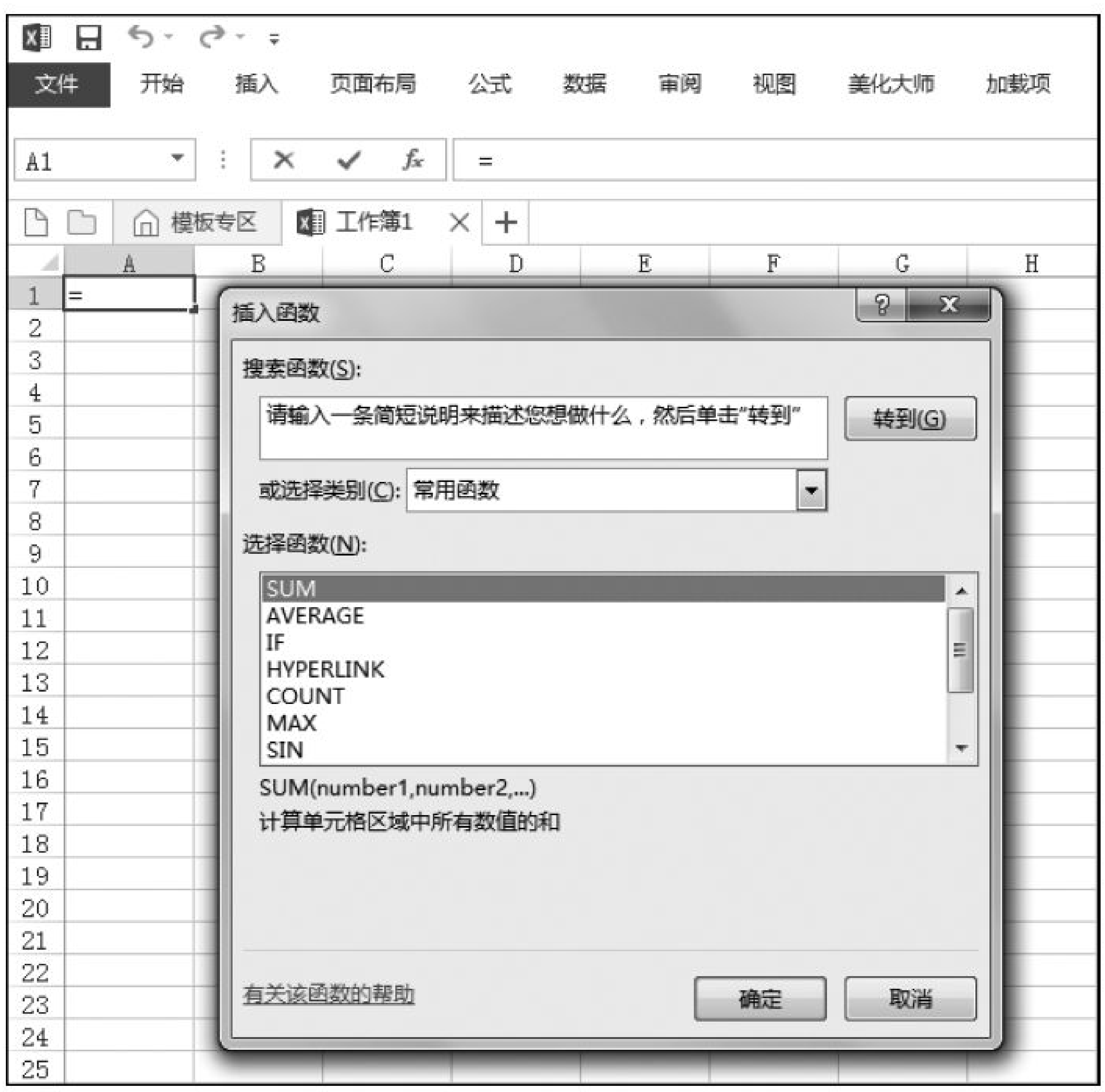


图 1-2 Excel 函数计算功能

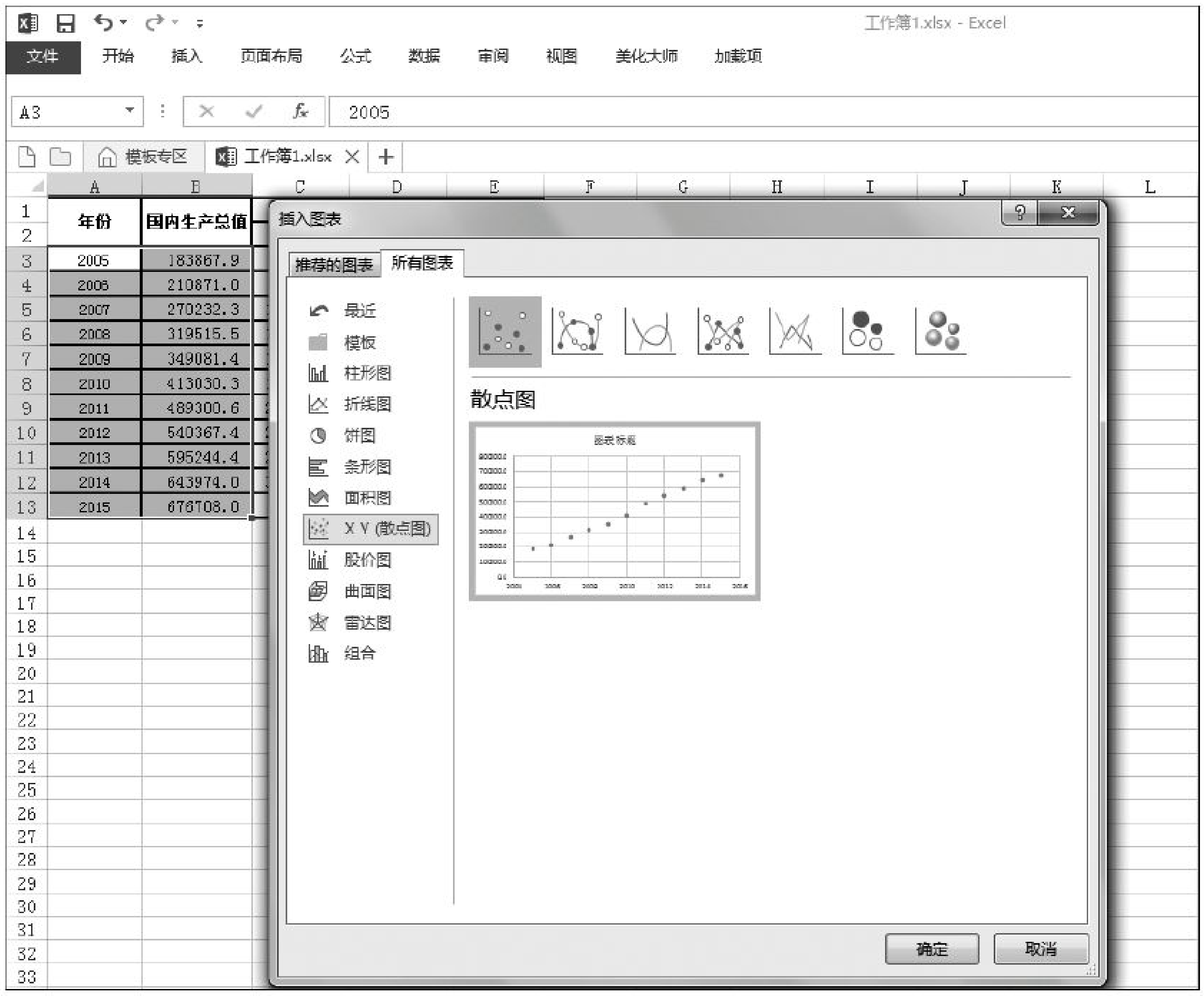


图 1-3　Excel 统计制图功能

Excel 提供了一组强大的数据分析工具，称为“分析工具库”。当需要进行复杂的统计或工程分析时，可以使用分析工具库节省步骤和时间(在较为简单的数据处理中很少涉及)。相对于使用函数输入公式求解复杂问题，使用分析工具库简易方便。

“分析工具库”并未直接安装显示在 Excel 功能区，需要启用加载项，并将加载项添加到功能区。安装的具体方法如下：

(1) 打开 Excel 之后，单击“文件”＞“选项”，在弹出的“Excel 选项”窗口中，依次单击“选项”＞“加载项”＞“分析工具库”，如图 1-4 所示。

(2) 单击“转到”按钮，弹出“加载宏”窗口，如图 1-5 所示。在该窗口中选择“分析工具库”选项，单击“确定”按钮。

(3) 单击“数据”选项卡，在最右侧就会出现“数据分析”按钮。

“分析工具库”安装完成后，当需要使用数据分析功能时，在 Excel 工作表的 “数据” 选项卡右侧的“分析”组中单击“数据分析”按钮，如图 1-6 所示，即可弹出“数据分析”对话框。

我们试图把你从复杂的计算中解脱出来，以便有足够的时间去理解统计方法的机理、把握统计思想，最终达到能够利用统计这一工具解决实际问题之目的。

我们坚信：读者如果能将统计方法与 Excel 应用软件结合起来，定会收到事半功倍的效果。

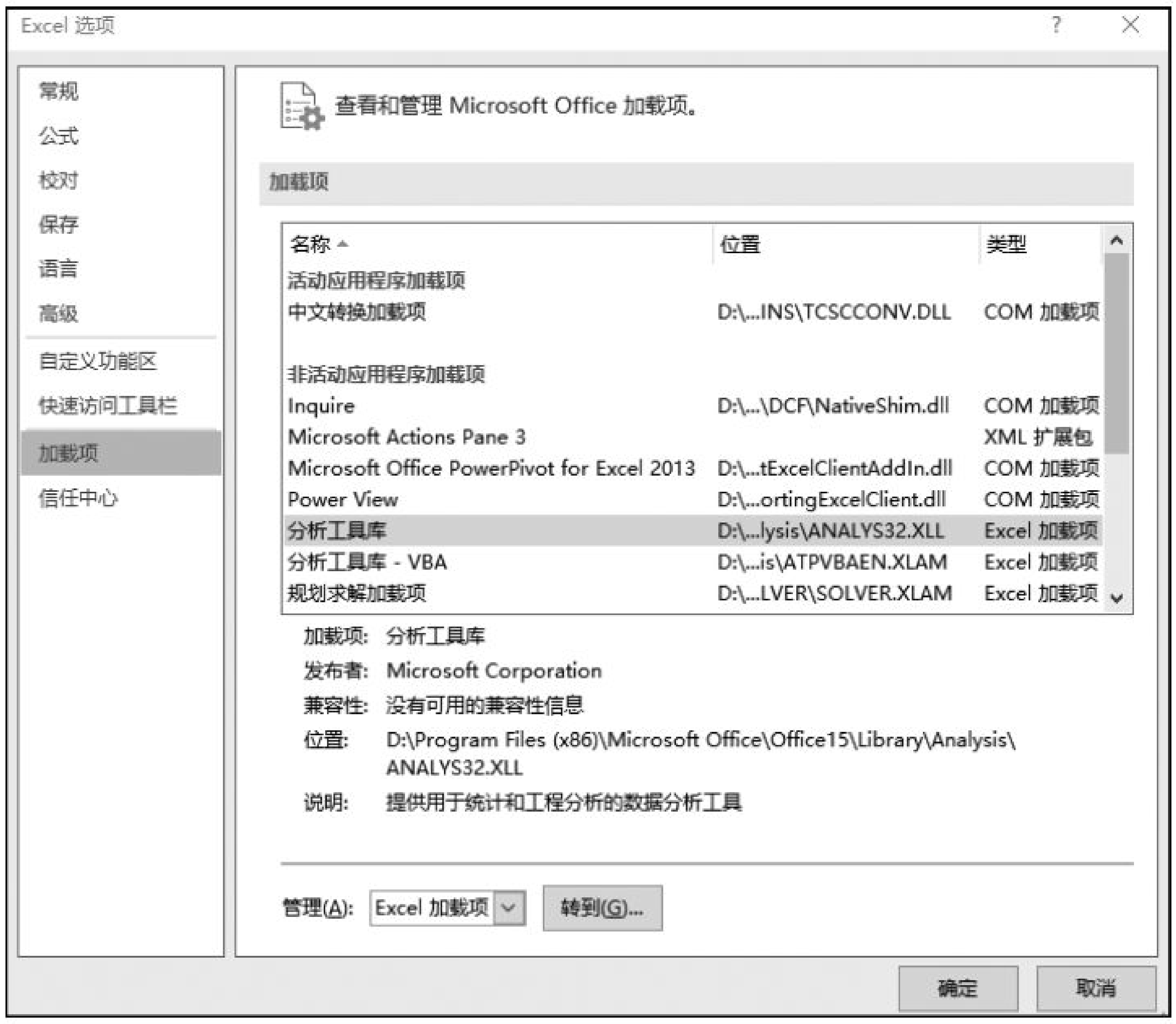


图 1-4 Excel 选项-加载项

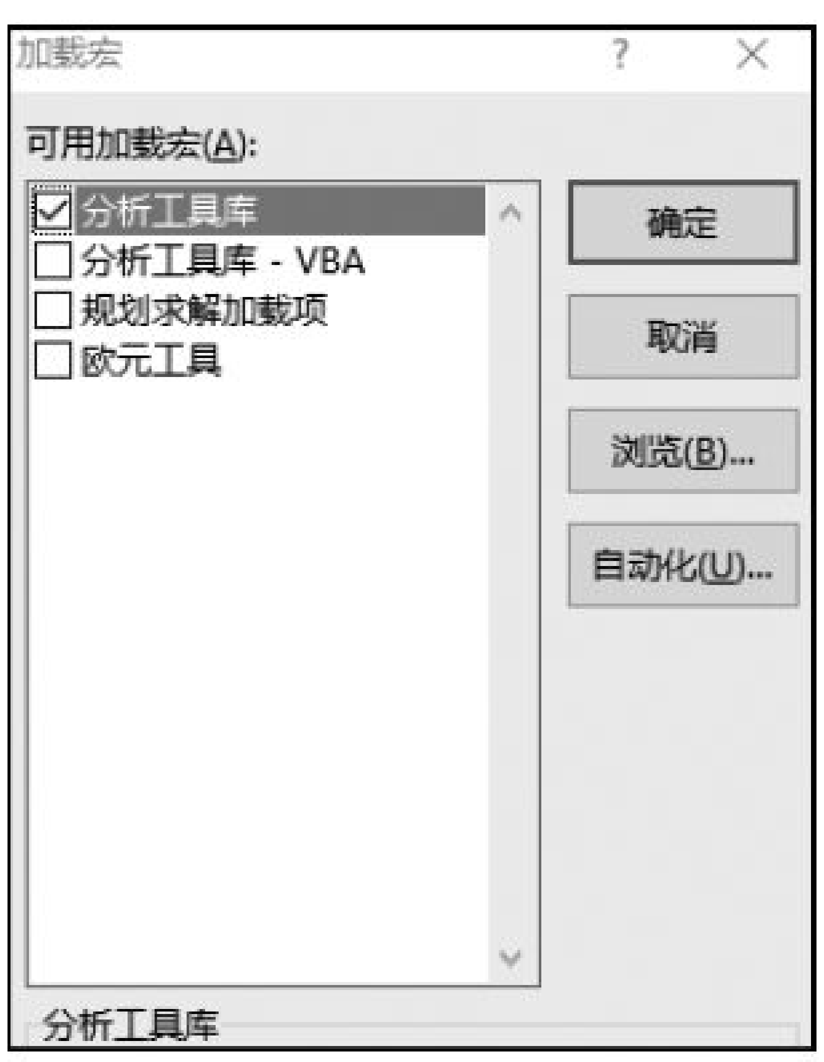


图 1-5 加载宏窗口



图 1-6 “数据分析”按钮

思 考 题

1. 简述统计一词的三个含义及其相互关系。
2. 简述统计学的研究对象。
3. 简述统计学的研究方法。
4. 简述统计的职能和任务。
5. 简述统计的研究过程。
6. 举例说明总体、总体单位、标志、指标之间的关系。

7. 区分四种计量尺度。
8. 归纳出统计法的主要内容。

练　习　题

2016 年某市统计部门提供的一份统计分析报告中有如下内容："我市国有工业企业 2 600 个，职工总人数 150 万人，工业总产值 360 亿元，人均实现产值 6 万元。其中，永达电气有限公司实现总产值 4 200 万元，职工人数 1 300 人。"

根据上述资料，回答下列问题：

(1) 指出该报告所反映的统计总体、总体单位分别是什么？

(2) 报告中涉及的统计标志有哪些？分别说明其性质。

(3) 报告中涉及的统计指标有哪些？

第二章

统计设计

【统计知识 ABC】

[A] 三驾马车

三驾马车的经济学意义指投资需求、消费需求、净出口(外部需求)三大需求,称之为拉动国民经济增长的"三驾马车"。统计学意义是指按支出法计算的国内生产总值中的最终消费、资本形成总额,以及货物和服务的净出口总额三大部分。

[B] 三次产业

三次产业是根据社会生产活动历史发展的顺序对产业结构的划分,产品直接取自自然界的部门称为第一产业,对初级产品进行再加工的部门称为第二产业,为生产和消费提供各种服务的部门称为第三产业。它是世界上较为通用的产业结构分类,但各国的划分不尽一致。

我国的三次产业划分如下。

第一产业:农业(包括种植业、林业、牧业和渔业)。

第二产业:工业(包括采掘业,制造业,电力、煤气及水的生产和供应业)和建筑业。

第三产业:除第一、第二产业以外的其他各业。由于第三产业包括的行业多、范围广,根据我国的实际情况,第三产业可分为两大部分:一是流通部门,二是服务部门。具体又可分为以下四个层次。

第一层次:流通部门,包括交通运输、仓储及邮电通信业,批发和零售贸易、餐饮业。

第二层次:为生产和生活服务的部门,包括金融、保险业,地质勘查业、水利管理业,房地产业,社会服务业,农、林、牧、渔服务业,交通运输辅助业,综合技术服务业,等等。

第三层次:为提高科学文化水平和居民素质服务的部门,包括教育、文化艺术及广播电影电视业,卫生、体育和社会福利业,科学研究业,等等。

第四层次:为社会公共需要服务的部门,包括国家机关、政党机关和社会团体以及军队、警察等。

(资料来源:http://www.stats.gov.cn/tjzd/tjzbjs/t20020327_14293.html.)

[C]三上企业

四大工程(2010 年 12 月,在全国统计工作会议上,国家统计局马建堂局长提出了四大工程的概念)包括:建设基本单位名录库、企业一套表制度、数据采集处理软件系统和联网直报系统,四大工程相互联系。

分步建设“三上”企业调查单位库和“三下”企业及非企业单位名录库，最终建成一个全国统一完整、不重不漏、真实准确、及时更新的基本单位名录库。

三上企业是指规模以上工业企业、限额以上批发零售住宿餐饮企业、资质以内的建筑业企业。其中：规模以上工业企业是指年主营业务收入在2 000万元以上的企业；限额以上批发企业是指主营业务收入2 000万元以上的企业，限额以上零售住宿餐饮企业是指主营业务收入在200万元以上的企业；资质以内的建筑业企业是指领取过建设部门颁发的《建筑业企业资质证书》（特级、一级、二级、三级和不分等级）的企业。

（资料来源：http://www.stats.gov.cn/tjzs/tjsdgc.pdf；http://fsh.bjstats.gov.cn/zsyd/27075.html.）

第一节　统计设计的含义

一、统计设计的概念

统计设计是统计工作的第一步，是指在做某项具体统计工作之前，根据统计研究目的结合研究现象（事物）的性质特点，对统计工作的各个方面、各个环节进行的打算（计划）和通盘考虑安排，制订的各种实施方案，如设计的指标、指标体系、分类目录、报表制度、调查方案、整理方案、分析方案等。

所谓各个方面，是指统计研究现象（事物）的各个组成部分，是统计工作的横向方面。对于微观企业统计来说，就是人、财、物、产、供、销等方面；对于宏观统计来说，就是社会、人口、经济、军事、政治、文化、教育、科技、卫生、体育、生产、分配、流通、消费、结构、比例、速度、效益等方面。

所谓各个环节，是指统计工作的各个阶段，是统计工作纵向方面。即统计设计、统计调查、统计整理、统计分析以及统计资料的提供、保存、公布等环节。

二、统计设计的分类

（一）整体设计和专项设计

从统计设计所包括的研究对象的范围看，统计设计分为整体设计和专项设计。

如果从统计研究对象的整体出发，对整个统计工作进行的全面设计就称为整体设计。如一个企业、一个省、一个国家整个统计工作的设计。

专项设计，是从统计研究对象的某一组成部分出发，对其某一项具体统计工作所做的设计。一个企业统计中的一个方面的设计，如人、财、物、产、供、销某一方面的统计设计，就属于专项设计。

（二）全阶段设计和单阶段设计

从统计设计所包括的统计研究工作阶段看，统计设计分为全阶段设计和单阶段设计。

全阶段设计是对统计工作各个阶段的全过程设计。从统计的目的、内容、指标到调查、整理、分析、资料提供、公布全过程的设计。

单阶段设计是指对统计工作全过程的某一阶段进行的设计。如仅对调查环节进行的

设计。

(三) 长期设计和短期设计

从统计设计所考虑的时间长短看,统计设计分为长期设计和短期设计。

长期设计是指对较长时期统计工作的设计,如"十二五"期间的统计工作设计。

短期设计是指对较短时期统计工作的设计,如 2012 年的统计工作设计。

三、统计设计的内容

(一) 明确统计研究目的

明确统计研究目的,是统计设计的首要环节。目的决定了研究对象和研究方法。根据统计研究目的并结合研究现象(事物)的性质特点,才能对统计工作的各个方面、各个环节进行打算(计划)和通盘考虑安排。

(二) 确定统计指标和指标体系

在明确统计研究目的的基础上,有针对性地、全面地从不同角度和方面设计出反映事物的统计指标与指标体系。

(三) 制订调查方案

制订调查方案就是制订如何去收集资料的方案。包括向谁调查、调查哪些项目、调查时间、调查组织形式、调查方法等。

(四) 制订统计整理方案

制订统计整理方案就是制订如何对调查数据进行整理的步骤和方法。包括数据的审核与插补;统计分组、分类;分布数列的编制;统计资料的显示方法等。

(五) 制订统计分析方案

制订统计分析方案就是制订如何进行统计分析的方案。包括选择分析的题目和方法;分析报告的撰写等。

(六) 安排各阶段工作进度

对各阶段所用的时间及整体进度进行安排。

(七) 协调各部门关系

协调统计工作所涉及的各部门之间的关系,保障统计工作顺利进行。

(八) 统计组织实施计划

统计组织实施计划包括机构的成立、人员的配备与培训、经费的筹措与使用等。

四、统计设计的原则

统计设计必须坚持以马克思主义、毛泽东思想、邓小平理论为指导，以科学发展观为统领，立足我国，放眼世界，吸收国内外先进的统计科研成果，按照客观规律，适应我国社会主义经济建设的需要。具体来说，进行统计设计应遵循下列原则。

（一）目的性

紧紧围绕统计研究目的和研究事物的特点进行统计设计。

（二）整体性

要站在一定高度上，整体把握，通盘考虑。

（三）客观性

统计设计一定要结合研究事物的特点，按照客观规律进行。

（四）具体性

具体性也就是可操作性。一项统计设计要具体、可操作、便于实施。

第二节　统计指标设计

一、统计指标的概念

指标（统计指标）是说明现象总体数量特征的概念或范畴。它有三大特点：数量性，所有统计指标都能用数量来表示；综合性，统计指标都是用来说明总体的；质的规定性，统计指标都是具有一定经济内容的。

统计指标，一般情况下是指指标名称或名称加数值。例如，2015 年中国国内生产总值（gross domestic product，GDP）676 708 亿元；全国财政收入达到152 217 亿元，比上年增长 8.4%都是统计指标。

统计指标的构成因素有：指标的名称、指标的概念、指标的内容、指标的计算方法、指标的计量单位、指标的空间范围、指标的时间范围、指标的数值八大因素。

二、统计指标的设计原则

设计一个统计指标，要遵循下列原则要求。

（1）指标结构（要素）要完整。包括指标名称、指标数值、时间范围、空间范围、计量单位。如 2015 年中国国内生产总值 676 708 亿元。当然在设计阶段，指标数值是没有的。提供指标时，就应该包括指标数值。

（2）指标的名称要有科学的理论依据，要做到名称和内容的统一。让人们看到名称就大体上知道指标所反映的内容。

居民消费者物价指数(consumer price index,CPI)反映居民生活中的产品和劳务价格所统计出来的物价变动指标,通常是作为观察通货膨胀水平的重要指标。

(3) 定义指标的概念。用确切的语言界定指标的内涵。

(4) 确定指标的外延或指标的内容。确定哪些现象数值计入,哪些不计入。

(5) 要有科学的计算方法和统计口径。

例如,工业总产值指标。工业总产值是以货币表现的工业企业在一定时期内生产的已出售或可供出售的工业产品总量,它反映一定时间内工业生产的总规模和总水平。它包括:在本企业内不再进行加工,经检验、包装入库(规定了需包装的产品除外)的成品价值,工业性作业价值,自制半成品、在产品期末初差额价值(生产周期较长的企业计算)。工业总产值采用"工厂法"计算,即以工业企业作为一个整体,按企业工业生产活动的最终成果来计算,企业内部不允许重复计算,不能把企业内部各个车间(分厂)生产的成果相加。但在企业之间、行业之间、地区之间存在着重复计算。

三、统计指标的分类

(一) 总量指标、相对指标和平均指标

统计指标按其所反映的内容或其数值表现形式,可以分为总量指标、相对指标和平均指标三种。

(1) 总量指标是反映总体规模的统计指标,通常以绝对数形式表现,故又称为绝对数,如人口总数、国内生产总值等。

(2) 相对指标是两个统计指标之比,称为相对数,如性别比例、第三产业所占比重、计划完成程度、经济增长率、股票价格指数等。

(3) 平均指标又称为平均数或均值,它所反映的是现象在某一空间或时间上的平均数量状况,如人均消费水平、某种股票一周平均价格、某班级统计学的平均成绩、单价、单产等。

广义上的相对数包括平均数。统计指标从表现形式上又分绝对数和相对数两种。所以统计分析一般要从绝对数和相对数两个方面进行。

(二) 数量指标和质量指标

统计指标按其所反映的数量特点或性质不同,分为数量指标和质量指标。

(1) 数量指标是指反映事物的规模大小或数量多少的统计指标,包括总体总规模、总水平或工作总量等方面的统计指标,如总人口、粮食总产量、工资总额、销售收入等。由于它反映的是事物总量,因此又称为总量指标。

(2) 质量指标是指反映事物数量关系、数量界限等质量方面的统计指标,包括总体内部数量关系、一般水平或工作质量,表明事物的结构、比例、速度、效益、强度、计划完成程度和一般水平。一般用相对数或平均数的形式表现。如人口性别比例、人口自然增长率、平均工资、居民消费价格指数等。

（三）实体指标和行为指标

统计指标按其所反映的客观实在和行为特征不同，分为实体指标与行为指标。

（1）实体指标是指反映客观存在的具有实物形态的统计指标，如钢铁产量、固定资产原值、劳动力人数等。

（2）行为指标是指反映某种行为数量的统计指标，如交通事故次数、犯罪率等。

（四）客观指标和主观指标

统计指标按其数据取值依据不同，分为客观指标和主观指标。

（1）客观指标是指其取值是通过实际度量或计数取得的，具有客观性，如产品产量、劳动生产率、GDP 等。

（2）主观指标是指凭借人们的感觉得到数据的统计指标，如民意调查当中的政府执政满意度等。

（五）描述指标、评价（考核）指标、监测（预警）指标

统计指标按其功能（作用）不同，分为描述指标、评价（考核）指标、监测（预警）指标。

（1）描述指标是指描述客观现象的数量方面（数量表现、数量关系和数量界限）的统计指标。如统计报表指标。

（2）评价（考核）指标是指对某一事物起到评价考核作用的统计指标，这些指标具有导向作用。如经济效益评价指标。

（3）监测（预警）指标是指对经济运行过程和结果进行追踪，看其是否偏离既定目标，起到预先警示作用的统计指标。如房地产价格调控指标、通货膨胀率等。

四、总量指标

（一）总量指标的概念

总量指标又称绝对指标或绝对数，是反映社会经济现象在一定时间、地点和条件下总规模或总水平的统计指标。总量指标通常是用绝对数来表现的，如 2010 年全国财政收入达到 83 080 亿元。

（二）总量指标的作用

总量指标是社会经济统计中最基本的统计综合指标，在实际工作中应用十分广泛。主要有以下几方面的作用。

（1）总量指标是对社会经济现象总体认识的起点。例如，一个国家的土地面积、国内生产总值、财政收入、粮食总产量、钢铁总产量等总量指标，反映了该国生产力发展水平和经济实力；一个地区的社会商品零售额、零售商业网点数等总量指标，标志着该地区的消费水平；一个企业的固定资产总值、工业增加值、商品产值、利税总额、职工人数等指标，则可反映该企业的人、财、物力的基本状况和生产经营活动成果。

(2) 总量指标是制定政策、编制计划、进行经济管理的重要依据。无论是宏观管理还是微观管理,都要从客观实际出发,而总量指标是最具体、最实际的客观数字。比如,制订国家经济发展战略和计划,就必须掌握国家的自然资源总量、人口数、主要产品产量等方面的总量指标,为编制计划和检查计划执行情况做好经济管理提供依据。

(3) 总量指标是计算相对指标和平均指标的基础。总量指标是基础指标,相对指标和平均指标都是由两个有联系的总量指标对比计算出来的,是总量指标的派生指标。因此,总量指标的计算是否科学、合理必将直接影响相对指标和平均指标的准确性。

(三) 总量指标的分类

按照不同的标准,可以将总量指标分为以下几种类型。

1. 按总量指标反映的内容不同,分为总体单位总量和总体标志总量

总体单位总量表示一个总体中所包含的总体单位总数,表示总体本身的规模大小,简称单位总量;总体标志总量是反映统计总体中各个单位某一数量标志值的总和,表示总体某一数量特征的总量,简称标志总量。就特定的统计总体来说,一定存在总体单位总量和总体标志总量两个基本的总量指标,两者相互联系、相辅相成。例如,研究某个工业企业职工的基本状况,该企业全体职工构成统计总体,职工人数是这个总体的单位总量,而工资总额就是这个统计总体的标志总量。一般来说,对于一个固定的总体,总体单位总量只有一个,而总体标志总量却有许多个。

2. 按总量指标反映社会经济现象的时间状况不同,分为时期指标与时点指标

时期指标是表明现象在一段时期内发展过程的总量指标。例如,一定时期的产品产量、工资总额、商品销售额等都是时期指标。时点指标是表明现象在某一时刻(瞬间)上发展状况的总量指标。例如,人口数、设备台数、商品库存量、储蓄存款余额等都是时点指标。

时期指标与时点指标相比各有不同特点。

(1) 时期指标的数值是连续计数的,即可以通过连续登记取得数据。它的每个数值都表示社会经济现象在这一时期内发生的总量;而时点指标的数值是间断计数的,即每隔一定时间登记一次,它的每个数值都表示社会经济现象发展到一定时点上所处的水平。

(2) 时期指标具有可加性,其相加累计的结果具有实际意义。例如,某企业 1 月产量是该月每天产量的总和,全年产量是 12 个月产量的总和。而时点指标不能累加,累加的结果没有独立的经济意义。例如,企业年末职工人数不等于各月末人数之和,各月末人数相加是没有意义的。

(3) 时期指标是流量指标,其数值大小与所属时期长短有直接关系,时期越长,数值越大,反之数值越小。例如,一个企业一年的销售额必然大于一个季度的销售额。时点指标是存量指标,其数值大小与时间长短没有直接关系。比如,企业年末职工人数不一定大于某月末的职工人数。

3. 按总量指标采用的计量单位不同,分为实物量指标、价值量指标、劳动量指标和统计分

(1) 实物量指标。

实物量指标是以实物单位计量的总量指标。实物量指标的最大特点是它直接反映现

象的使用价值和经济内容，能够具体表明事物的规模或水平。因而广泛用来反映国民经济的基本情况、制订和检查计划、研究各种产品对生产和生活的满足程度等。

实物单位，是根据事物的自然属性和特点而采用的计量单位。主要有：自然单位、度量衡单位、标准实物单位、复合单位和双重或多重单位等。

自然单位，是指按照被研究现象的自然状态计量的单位，通常用于可数现象的计量。如人口按“人”、汽车按“辆”、鞋子按“双”、设备按“台”为单位。

度量衡单位，是指根据国内或国际上通行的度量衡制度对现象进行计量的单位，通常用于不可数现象的计量。如粮食以“千克”或“吨”为单位，布匹长度用“米”为单位，天然气以“立方米”为计量单位等。

标准实物单位，是按照统一的折算标准来度量被研究现象数量的一种计量单位。在很多情况下，同类现象之间往往存在大小或品质方面的差异，如拖拉机的牵引能力、煤的发热量等，对这些现象不加区别、直接汇总，就不能客观反映这些现象的实际使用价值，所以必须对其按规定的标准进行计量。如拖拉机折合标准台，不同含量的化肥折合为100%的含量计算，各种不同发热量的煤折合为7 000大卡[①]/千克的标准煤等。

复合单位，是将两种计量单位结合在一起对研究现象进行计量的单位。当用一种计量单位无法准确地反映现象的总规模和水平时，往往采用复合单位进行计量。如运输企业核算货运量，只用载重量或运输距离均不能准确反映运输结果，故采用载重量乘运输距离的方法，以“吨千米”为计量单位。又比如，发电量用“千瓦时”表示，等等。

双重或多重单位，是同时采用两种或两种以上计量单位对被研究现象进行计量的单位。例如，发动机的数量可以用台，也可以用千瓦表示，同时用台和千瓦表示即“台/千瓦”就是双重计量单位。又如船舶同时用艘、马力、吨位三种单位表示，就属于多重计量单位。复合单位、双重或多重单位可以更有效地反映出现象的使用价值或经济内容。

实物量指标又进一步分为混合实物量和标准实物量。

混合实物量是指把名称相同、用途相同但规格或成分不同的同类事物的实物量直接加总所得到的总量。如产品的混合量可以概括地粗略地说明某类产品的使用价值量，但不确切。

标准实物量是指把不同规格或成分的同类事物的实物量，按照一定的折算系数折算成某一标准规格或含量的实物总量。其计算公式为

$$\text{标准实物量} = \sum(\text{实物量} \times \text{折算系数})$$

$$\text{折算系数} = \frac{\text{实际规格或含量}}{\text{标准规格或含量}}$$

例如，某企业6月生产32支纱100吨；生产18支纱100吨；生产54支纱100吨。则该企业6月纺纱混合产量为100＋100＋100＝300(吨)；标准产量(32支纱)为100＋100×18/32＋100×54/32＝325(吨)。

(2) 价值量指标。

价值量指标是以货币为单位来计量的总量指标，反映社会经济现象价值量，如国内生

---

① 1大卡＝4 185.851 820 846焦耳。

产总值、工业增加值、销售收入、产品成本等。

与实物量指标相比,价值量指标具有最广泛的综合性和较强的概括力。它可以把那些不能直接加总的产品或商品数量过渡到可以直接加总,进而综合说明不同使用价值的产品或商品的总水平、总规模或总速度等。但是,价值量指标脱离了具体的物质内容,比较抽象,不能确切地反映实际情况。因此,价值指标应该与实物指标结合使用,以便全面说明问题。由于价格和价值时常背离,如果用价值量指标反映生产发展速度,必须剔除价格变动的影响。

(3) 劳动量指标。

劳动量指标是以劳动单位为计量单位的总量指标。劳动单位是由劳动和工作时间组合而成的复合单位,如"工时""工日"等。如一个工人工作一小时称为一个工时,工作一天称为一个工日。例如,某车间有 50 个工人,生产某种产品各工作了 8 小时,则该车间共工作了 400 个工时。对于多种产品产量,可用定额工时来计量,表现为定额工时产量。其计算公式为

$$\text{定额工时产量} = \sum(\text{工时定额} \times \text{合格产品产量})$$

例如,某车间在一天内生产了甲、乙、丙、丁、戊五种产品,其产量分别为 24、12、40、4、8,其单位产品工时定额分别为 0.4、0.5、0.8、1.5、1.0,则该车间完成的定额工时产量为 61.6(24×0.4+12×0.5+40×0.8+4×1.5+8×1.0)。

以劳动单位计算的劳动总消耗量指标,主要用于企业内部管理,反映劳动总成果,评价劳动时间利用程度和计算劳动生产率,等等,可用于劳务工资的核算,也可以用来编制生产作业计划并检查生产计划执行情况等。劳动量指标在基层生产企业中应用最普遍。

(4) 统计分。

有些行业或项目统计,往往以统计分作为统计计量单位。如产品质量打分;体育行业的体操项目打分、篮球项目计分;文娱行业演唱比赛打分;学生考试成绩得分等。

统计分还可进一步分为一般统计分和换算标准分。标准分的换算见平均分析一章中的差异分析一节。

(四) 总量指标的计算

1. 总量指标的计算方法

总量指标的计算方法,就是对所研究的现象进行计数、测量的方法。它是大量个体绝对数据汇总的结果。有简单加法和加权加法两种。其计算公式分别为

简单加法:$M = x_1 + x_2 + \cdots + x_n = \sum x$

加权加法:$M = x_1 f_1 + x_2 f_2 + \cdots + x_n f_n = \sum xf$

式中,M 为总量指标;x 为变量值;f 为权数。

在未分组资料条件下,用简单加法公式;在分组资料条件下,用加权加法公式。标准实物量和定额工时产量的计算公式可归并到加权加法公式之中。

2. 总量指标的计算要求

(1) 必须注意现象的同质性。同质性是由现象的性质和用途决定的,不同性质的事

物是不能直接计算其总量的。但是，对一些具体形式不同但使用价值相同的产品，可以折算为标准品产量。如原煤、柴油、汽油、天然气、电等产品，可以折算成吨标准煤产量。

（2）明确总量指标的含义、范围和计算方法。任何一个总量指标，都是一定经济范畴的具体数量表现，有一定质的规定性。它的计算不是单纯的汇总问题，必须对总量指标的含义、包括的范围和计算方法做出严谨的、科学的界定后才能计算。例如，在计算工业总产值指标时，首先要明确什么是工业和工业产品，否则，就不可能准确地对工业总产值指标进行统计。再比如，国内生产总值可以采用支出法、生产法和收入法三种计算方法，计算方法不同，结果就不一样。因此，一定要根据研究目的，统一规定指标的含义，确定科学的统计方法。

（3）要使用统一的计量单位。为便于进行统计比较和综合分析，同一总量指标在不同时间、地点和单位进行统计时，计量单位应当一致，如果存在不一致的情况，应采用一定的方法进行换算。

五、相对指标

（一）相对指标的概念

相对指标又称相对数，是两个有联系的统计指标的比值，用以说明社会经济现象之间的数量对比关系。用来对比的两个数，既可以是绝对数，也可以是平均数和相对数。比如，2006 年我国国内生产总值指数为 111.1%、人口自然增长率为 5.28‰、居民消费价格指数为 101.5%，城镇居民家庭平均每百户拥有移动电话 153 部等，都是相对指标。

（二）相对指标的作用

相对指标的作用有以下两点。

（1）相对指标可以综合反映现象的相互关系、内部结构、实现的程度、强度和速度等，从而说明总量指标不能充分表达的问题。社会经济现象之间是相互联系、相互依存的，不论是从时间上、空间上还是事物内部结构上都是如此。因此，要分析某一社会经济现象，仅利用总量指标是远远不够的，还必须计算相对指标，全面地反映现象总体的内部结构、比例关系，现象的普遍程度、普及程度和发展速度。同总量指标相比，相对指标揭示了现象内涵方面的特征，它对现象总体数量特征的描述和反映是动态的、深入的。

（2）相对指标可以使某些利用总量指标不能直接进行对比的社会经济现象，取得可以比较的基础。总量指标数值容易受总体范围的影响，总体范围越大，指标数值也就越大，反之就越小。为了消除由于总体范围不同给统计对比分析带来的影响，就要使用相对指标。例如，某企业甲班组某月生产时出现废品 60 件，乙班组出现废品 50 件，我们不能简单地认为甲班组的质量比乙班组差。因为还要考虑两个班组生产产品总量和劳动效率等因素。如果甲班组的总产量为 1 200 件，乙班组的总产量为 800 件，则甲班组的废品率为 5%（60÷1 200），乙班组的废品率为 6.25%（50÷800），利用相对指标可以确切地表明甲班组的废品率要低于乙班组，显然，甲班组的质量好。

(三) 相对指标的表现形式

相对指标的数值,通常有两种表现形式:有名数和无名数。

1. 有名数表示的相对指标

有名数表示的相对指标,主要用以表现事物的强度、密度和普遍程度。将分子指标和分母指标的计算单位同时使用,即采用复合单位计量。如人口密度用“人/平方千米”表示,城市人口拥有公共汽车用“辆/万人”表示,人均粮食产量用“千克/人”表示等。

2. 无名数表示的相对指标

无名数是一种抽象化的数值,常以系数、倍数、成数、百分数、千分数或翻番数等表示。

(1) 系数和倍数。系数和倍数是将对比基数抽象化为 1 而计算的相对数。当分子数值和分母数值比较接近时,常用系数表示,如果子项特别大而母项特别小,则常用倍数。

(2) 成数。成数是将对比的基数抽象化为 10 而计算出来的相对数。如某地区 2017 年粮食产量比 2016 年增长两成,即增产 2/10。

(3) 百分数。百分数是将对比的基数抽象化为 100 而计算的相对数。

(4) 千分数。千分数是将对比的基数抽象化为 1 000 而计算的相对数,适用于分母数值特别大而分子特别小的情形。如人口出生率、死亡率、自然增长率等。

(5) 翻番数。翻番数是指两个相比较的数值中,一个数是另一个数的“2^m”倍,其中,m 是番数。例如,党的十七大报告提出全面建设小康社会新的更高要求,“实现人均国内生产总值到 2020 年比 2000 年翻两番”。2000 年我国人均国内生产总值为 7 858 亿元,则 2020 年的人均国内生产总值应达到 31 432($7\ 858\times2^2$)亿元。

(四) 相对指标的种类和计算方法

在社会经济统计中,常用的相对数指标可分为计划完成相对数、结构相对数、比例相对数、比较相对数、动态相对数、强度相对数和平均数七种。

1. 计划完成相对数

计划完成相对数也称计划完成百分比,它是现象在某一时期的实际完成数与其计划任务数的比值,用来检查、监督计划的执行情况,通常用百分数(%)的形式来表示。其基本计算公式为

$$\text{计划完成相对数}=\frac{\text{实际完成数}}{\text{同期计划数}}\times100\%$$

式中,子项实际完成数是根据实际完成情况进行统计而得到的数据,母项则是下达的计划指标,子项和母项在指标含义、计算方法、计量单位以及时间长度等方面应完全一致。同时,分子、分母不允许互换。

实际工作中,由于计划数可表现为绝对数(总量指标)、相对数(相对指标)、平均数(平均指标)等多种形式,因此计算计划完成相对数的方法也不尽相同。

(1) 计划数为绝对数。

当计划任务数为绝对数即总量指标时,它一般适合于考核社会经济现象的规模或水平的计划完成情况。

① 短期计划检查。可直接用实际完成数与计划完成数对比求得计划完成程度指标。计算公式为

$$计划完成相对数=\frac{实际完成数}{计划完成数}\times100\%$$

② 中长期计划检查。在检查中长期计划(如五年计划)任务的完成情况时,根据计划指标的性质不同,分为水平法和累计法两种。

水平法,适用于反映生产能力的经济指标,如钢产量、煤产量、发电量、粮食产量等指标的计划完成情况检查。它是以计划期末(最后一年)应达到的水平为对象考核的。计算公式为

$$计划完成相对数=\frac{计划期末(最后一年)实际达到的水平}{计划规定期末应达到的水平}\times100\%$$

按水平法计算提前完成计划的时间,是按照连续一年(可以不在一个自然年度,只要是连续12个月即可)的实际数值,与长期计划规定最后一年的计划任务数对比,若二者相等,说明完成了长期计划,剩余的时间即为提前完成长期计划的时间。

例如,“十二五”计划规定在最后一年(2015年)某种产品产量达到200万辆,实际为220万辆,则“十二五”计划完成程度为:220/200=110%。

假定从2014年7月至2015年6月该产品产量刚好达到了200万辆,则提前了半年完成了“十二五”计划。

累计法,它是以计划期内各年计划数量的累计总和为对象考核的。适用于检查计划期内构成国民财产存量的经济指标,如基本建设投资额、造林面积、住宅建设、开垦荒地等计划完成情况。计算公式为

$$计划完成相对数=\frac{计划期间实际累计完成数}{计划期间规定的累计数}\times100\%$$

按累计法计算提前完成计划的时间,是用整个计划期的时间减去自计划执行之日起到累计实际完成数刚好达到计划任务数为止的时间求得。

例如,“十二五”计划规定基本建设投资额500亿元,实际完成了550亿元。则“十二五”计划完成程度为:550/500=110%。

假定从2011年1月1日起,到2015年6月30日累计实际完成基本建设投资额达到了500亿元,则提前了半年完成了“十二五”计划。

③ 进度计划检查。在检查计划执行情况时,为了分析计划执行进度,预计计划执行结果,在期中需计算计划执行进度指标。其计算公式为

$$进度计划完成程度=\frac{自报告期初止累计完成数}{全期计划任务数}\times100\%$$

例如,某公司全年计划实现利润1 000万元,上半年实际完成了600万元,则该公司上半年就完成了全年计划的60%(600/1 000)。

计划执行进度要和时间流逝相一致,即时间过半,任务也要过半,否则全期计划的完成就没有了时间的保证,应及时分析其原因,以便提出可行的对策,确保计划的完成。

(2) 计划数为相对数。

当计划任务数为相对数,而且以提高(或降低)百分数表示时,则不能以实际提高(或

降低)百分数与计划提高(或降低)百分数直接对比,应在原有基数 100%的基础上提高(或降低),才能对比。其计算公式为

$$计划完成相对数=\frac{实际达到的百分数}{计划规定的百分数}\times 100\%$$

$$=\frac{1\pm 实际提高(降低)百分数}{1\pm 计划提高(降低)百分数}$$

例如,某企业产量计划提高 10%,实际提高 15%,则

$$产量计划完成相对数=\frac{1+15\%}{1+10\%}\times 100\%=104.55\%$$

计算结果表明,该企业产量超额 4.55% 完成计划。

若计划为相对数,可以用相减的方法来检查计划完成情况。如上例,15%-10%=5%,计算结果表明,该企业产量比计划提高 5 个百分点。百分点是百分数中相当于 1%的单位,它与百分数不同。专用于表示同一内容在不同时期或不同地点的指标百分数的变化。

(3) 计划数为平均数。

当计划任务数为平均数时,计划完成程度的计算公式为

$$计划完成相对数=\frac{实际完成的平均数}{计划规定的平均数}\times 100\%$$

例如,某企业计划要求劳动生产率达到 5 000 元/人,某种产品单位成本计划为 100 元。该企业实际劳动生产率达到 6 000 元/人,实际单位成本为 90 元,则计划完成程度指标为

$$劳动生产率计划完成相对数=\frac{6\ 000}{5\ 000}\times 100\%=120\%$$

$$单位成本计划完成相对数=\frac{90}{100}\times 100\%=90\%$$

计算结果表明,该企业劳动生产率超额 20%完成了计划任务,单位成本也超额 10%完成了计划任务。

需要说明的是,在分析计划完成情况时,要注意计划任务数的性质差异。凡是计划指标是以最低限额规定的,如产品产量、利润、销售额等指标,计划完成相对数等于或大于 100%,才算完成或超额完成计划。反之,凡是计划完成相对数是以最高限额规定的,如原材料消耗、单位成本、流通费等,其计划完成相对数等于或小于 100%,才算完成或超额完成计划。

2. 结构相对数

结构相对数是表明总体内部各个组成部分在总体中所占比重的相对指标,也称比重相对数,一般用百分数表示。结构相对数是描述总体特征的重要指标,它可以说明总体内部构成,是分析总体分布的基础。计算公式为

$$结构相对数=\frac{总体中某一部分数值}{总体全部数值}\times 100\%$$

由于结构相对数是根据同一总体资料计算的,其分子只能是总体的一部分,因此个别结构相对数是一个大于 0 小于 1 的数值,而总体中各部分比重之和等于 100%或 1。

结构相对数是描述总体特征的重要指标,它可以说明总体内部构成,分析总体内部构成的变化,反映事物发展变化的过程及趋势。如第三产业结构及其变化。

3. 比例相对数

比例相对数是反映一个统计总体内部各个组成部分之间数量对比关系的相对指标,常用系数和倍数表示。

$$比例相对数=\frac{总体中某一部分数值}{总体中另一部分数值}$$

例如,2015 年我国男性人口为 70 414 万人,占总人口的 51.22%;女性人口为 67 048 万人,占总人口的 48.78%,则人口性别比为 105.02。

比例相对数的计算基数单位通常为 1、100、1 000 等。例如,我国计算农民和非农民的消费比例关系就是以农民消费水平为 1 计算的。在比例相对数的计算过程中,子项和母项可以互相颠倒,当然,分子和分母颠倒后,指标数值所说明的问题也正好反过来了。

比例相对数也有反映总体结构的作用,与结构相对数有着密切的联系,它是属于一种结构性的比例,两者作用相同,只是使用的方法不同,侧重点有所区别。计算比例相对指标对于分析研究国民经济的平衡比例关系、保持国民经济稳定协调发展具有重要意义。国民经济中客观存在着各种各样的比例关系,如物质生产部门和非物质生产部门的关系;积累和消费的比例关系;社会再生产中生产、流通、分配的比例关系;等等。

4. 比较相对数

比较相对数是反映同一时期的同类现象在不同地区、部门和单位之间数量对比关系的相对指标。其结果可以用百分数表示,也可以用倍数表示。计算公式为

$$比较相对数=\frac{某地区(单位)某期某类指标数值}{另一地区(单位)同期同类指标数值}$$

例如,2015 年美国 GDP 为 18 124.73 十亿美元,中国为 11 211.93 十亿美元。则

2015 年美国与中国的 GDP 比较相对数指标=18 124.73÷11 211.93=1.617。

计算结果说明,2015 年美国 GDP 是中国的 1.617 倍。

运用比较相对数,可以揭示同类现象在不同空间条件下的差异程度。根据研究目的和方式不同,比较相对指标的分子指标和分母指标可以互换位置。它既可用于不同国家、地区、部门或单位之间的比较,也可用于先进与落后的比较,还可用于和标准水平或平均水平的比较。通过对比可以找出差距和不足,为管理决策提供参考依据。

计算比较相对数指标,可以用绝对数,也可以用相对数或平均数,但不管用何种指标计算,都必须考虑是否具有可比性。

5. 动态相对数

动态相对数又称发展速度,它是把同一现象在不同时间上的指标数值进行对比的比值,用以说明现象发展变化的方向和程度。通常用百分数表示,也可用倍数或翻番表示。其计算公式为

$$动态相对数=\frac{某一现象报告期数值}{同一现象基期数值}\times 100\%$$

例如,中国粮食总产量 2015 年为 62 143.5 万吨,2014 年为 60 702.7 万吨,则我国粮食产量的发展速度为 62 143.5/60 702.7≈102.4%。

动态相对数指标的表现形式有多种,本书将在第六章时间序列分析中详细介绍。

6. **强度相对数**

强度相对数是两个性质不同但又相互联系的总量指标进行对比的比值,可以反映现象的强度、密度或普遍程度。计算公式为

$$强度相对数=\frac{某一总量指标的数值}{另一有联系但性质不同的总量指标数值}$$

强度相对数的表现形式一般是复名数,由分子指标和分母指标的计算单位组成。如人均国民生产总值“元/人”,人口密度“人/平方千米”等。有的强度相对数用次数、倍数、系数、百分数或千分数表示,如高炉利用系数、资金周转次数、流通费用率、人口出生率等。

例如,2015 年我国人口总数约为 137 462 万人,国土面积约为 960 万平方千米。则:我国人口密度=137 462/960=143.189 6(人/平方千米)。

大多数用复名数为单位的强度相对数指标,其分子指标和分母指标可以互换,因此,派生出正指标和逆指标两种形式。正指标是指标数值大小与现象的发展程度或密度成正向变化的强度相对数指标,即指标数值越大,现象的发展程度或密度越高,反之就越低;逆指标是数值大小与现象的发展程度或密度成反向变化的强度相对数指标,即指标数值越大,现象的发展程度或密度越低,反之就越高。

例如,某地区 2015 年总人口为 1 200 万,有 60 000 个零售商业机构,则该地区零售商业网点密度指标是多少?

零售商业网点密度指标是衡量一个国家或地区商业发展水平的统计综合指标,可以用两种方法计算。

$$正指标:零售商业网点密度=\frac{某地区零售商业机构数}{该地区人口数}=\frac{60\ 000}{1\ 200}=50(个/万人)$$

即平均每万人拥有 50 个零售商业机构为他们服务。

$$逆指标:零售商业网点密度=\frac{某地区人口数}{该地区零售商业机构数}=\frac{12\ 000\ 000}{60\ 000}=200(人/个)$$

即每个零售商业机构为 200 人服务。

正指标越大,说明商业网点密度越高;反之,就越低。逆指标越大,说明商业网点密度越低;反之,就越高。

需要指出,计算强度相对指标必须注意社会经济现象之间客观上要存在一定的经济或技术上的联系,这样,两个指标对比才会有现实意义。

7. **平均数**

平均数又称平均指标,是同质总体内某一数量标志在一定时间、地点条件下达到的一般水平,即总体内各单位参差不齐的标志值的代表值。平均指标的特点是对数量标志在总体单位之间的数值差异抽象化,它并不代表某一总体单位的具体水平,而是一个用来反映总体综合数量特征的典型水平或者代表水平。例如,用职工平均工资代表职工收入的一般水平,用平均亩产代表粮食生产的一般水平等。

平均数的基本计算公式为

$$平均数(算术平均)=\frac{总体标志总量}{总体单位总量}$$

例如,职工平均工资=工资总额/职工人数,平均亩产量=粮食总产量/播种面积。

(五) 计算和运用相对指标的原则

相对指标种类较多,各有自己的意义和作用,在计算和应用上也有一些不同的要求与特点,但从总体上说,计算和应用相对指标,要坚持四个方面的原则。

(1) 可比性原则。遵从可比性是计算相对指标应把握的总原则,两个对比的指标是否具有可比性,是计算结果能否正确反映现象之间数量对比关系的重要条件。可比性是指相互对比的两个统计指标,在经济内容、计算范围、计算方法和计量单位等方面保持一致,如果不可比,就需要进行调整。

(2) 正确选择对比基数的原则。基数是相对指标对比的依据或标准,基数选择不当,就会失去相对指标的作用,对同一现象、同一问题,采用不同的对比基数就会有不同的结果。因此,基数的选择必须从统计研究的目的和任务出发,结合研究对象的性质和特点,选择能最大限度反映现象内在联系和本质特征的指标作为对比基数。

(3) 多种相对指标结合运用的原则。因为客观经济现象是复杂的,而每一种相对指标只是反映经济现象的某一个侧面。在分析研究复杂的现象时,应该将多种相对指标结合起来使用,多角度、多侧面地分析研究经济现象,有助于深入、全面地分析问题和认识问题。

(4) 相对指标与总量指标结合运用的原则。总量指标能够反映事物发展的总规模和总水平,却不易看清事物差别的程度;相对指标反映了现象之间的数量对比关系和差异程度,却又将现象的具体规模和水平抽象化了,而无法说明现象之间的绝对量差异。因此,只有将相对指标同总量指标结合起来使用,才能克服认识上的片面性,实现对客观事物全面、正确的认识。

第三节　统计指标体系设计

一、统计指标体系的含义

统计指标体系是指由若干相互联系的统计指标构成的有机整体。一个统计指标仅仅从一个侧面反映了社会经济现象,而社会经济现象是多方面的、复杂的,要全面、系统地反映社会经济现象,就必须建立统计指标体系。

例如,盲人摸象的寓言故事,告诉我们看问题要全面,防止片面性的马克思主义哲学思想,就是建立指标体系的理论基础。机械制图中用到的三视图,说明对于一个立体物体,至少要从三个方面(主视、左视和俯视)进行观察或反映。“横看成岭侧成峰,远近高低各不同。不识庐山真面目,只缘身在此山中”的诗句,也折射出建立指标体系的缘由。

二、统计指标体系的设计原则

设计指标体系的基本要求如下。

(1) 科学性,即指标体系的设计要符合事物的特点,每一指标的设计都符合统计指标

的设计要求。

(2) 目的性,即指标体系的设计要考虑管理的要求或研究目的。

(3) 全面性,即指标体系的设计要全面,从不同侧面反映事物。

(4) 统一性,即指标体系的设计要三大核算统一。

(5) 可比性,即指标体系的设计要不同空间、不同时期可比。

(6) 核心性,即指标体系的设计要确定核心指标。

(7) 可行性,即指标体系的设计要保证每一指标都能取得。

(8) 互斥性,即指标体系的设计要使指标之间相关程度弱,注重指标的代表性。

三、统计指标体系的形式

社会经济总体现象的构成往往是复杂多样的,其间的联系形式是多种多样的,因而构成统计指标体系的各项指标之间的联系也有多种不同的形式,归纳起来主要有两种。

一类是内容衔接结构式统计指标体系,是用一系列统计指标所构成的从不同侧面、不同角度、不同环节、不同层次来反映研究对象不同数量特征的指标体系。

例如,反映一个企业从人、财、物、产、供、销等方面设计的指标体系。

另一类是恒等式数量关系指标体系,是由三个及三个以上的统计指标所形成的具有一定数量对等关系的指标体系。这类指标是建立统计指数体系的基础,也是进行因素分析的基本依据。

例如,原材料消耗额、产品产量、单位产品原材料消耗量、单位原材料价格。

四、统计指标体系的种类

(一) 描述性指标体系、评价性指标体系、监测性指标体系

按统计指标体系的功能层次来分,有描述性统计指标体系、评价性统计指标体系、监测性统计指标体系三种。其中,描述性统计指标体系是基础,它的任务是尽量全面、详细描述研究现象总体各方面数量特征,因此指标数量多,指标形式不一,指标之间重叠交叉现象普遍。评价性统计指标体系则是在描述性统计指标体系中根据评价任务精选出来的若干指标,或称“指标样本”,其指标数量相对少而精,在不失全面的情况下突出重点,指标形式一般比较统一,指标之间重叠交叉现象一般尽量避免。监测性或决策性指标体系则更加精练,一般是由少数几个精干的指标构成,甚至于只由一些高度综合的指标构成。

(二) 社会统计指标体系、经济统计指标体系、科技统计指标体系

按指标体系反映的内容不同分为社会统计指标体系、经济统计指标体系和科技统计指标体系。

(三) 宏观统计指标体系和微观统计指标体系

按指标体系反映现象的范围不同分为宏观统计指标体系和微观统计指标体系。

（四）基本统计指标体系和专题统计指标体系

按指标体系反映现象的方面侧重点不同分为基本统计指标体系和专题统计指标体系。

基本统计指标体系是由反映社会经济发展基本情况的统计指标构成的。专题统计指标体系是针对某一专门问题而设计的指标体系，如经济效益指标体系。

五、经济效益指标体系

（一）经济效益的含义

经济效益是由多种因素构成的一个复杂的经济范畴。与之相关联的还有两个范畴：经济效果和经济效用。

经济效果是某一具体经济实践过程的产出与投入之比，即是该过程的有用成果与劳动消耗之比。一定量的投入，包括物化劳动和活劳动投入，取得的有用成果越大表明经济效果越好。或者说一定量的产出，即生产一定量的有用成果，所耗的劳动越少表明经济效果越好。用公式表示为

$$\text{经济效果}=\frac{\text{有用成果}}{\text{劳动消耗}}\quad\text{或}\quad\text{经济效果}=\frac{\text{劳动消耗}}{\text{有用成果}}$$

经济效用是满足人们需要的量与有用成果之比。人们从事各种经济活动，生产各种产品或提供各种劳务，其目的是满足人们的需要。商品或劳务满足人们需要的能力在经济学上称为效用。一定量的有用成果，满足人们需要的程度越高表明经济效用越大。
用公式表示为

$$\text{经济效用}=\frac{\text{满足需要}}{\text{有用成果}}\quad\text{或}\quad\text{经济效用}=\frac{\text{有用成果}}{\text{满足需要}}$$

经济效益则是经济过程取得的经济效果及其带来的经济效用的总称，即是满足人们需要的量与劳动消耗之比。用公式表示为

$$\begin{aligned}\text{经济效益}&=\frac{\text{满足需要}}{\text{劳动消耗}}=\frac{\text{有用成果}}{\text{劳动消耗}}\times\frac{\text{满足需要}}{\text{有用成果}}\\&=\text{经济效果}\times\text{经济效用}\end{aligned}$$

一定量的投入，取得的有用成果越大，同时满足人们需要的程度越高，表明经济效益越好。满足需要包括生活需要和生产适销对路的产品，满足需要的程度高，经济效用大，才可能给企业带来良好的经济效益。反之若生产的产品适销不对路，尽管也算是有用成果，有经济效果，但不能满足人们的需要，经济效用小，也就谈不上经济效益。

（二）国家统计局颁布实施的工业经济效益评价考核指标体系

1. 工业经济效益考核指标体系内容及计算公式

(1) 总资产贡献率。该指标反映企业全部资产的获利能力，是企业管理水平和经营业绩的集中体现，是评价和考核企业盈利能力的核心指标。

$$\text{总资产贡献率}=\frac{(\text{利润总额}+\text{税金总额}+\text{利息支出})}{\text{平均资产总额}}\times 100\%$$

其中：税金总额为产品销售税金及附加与应交增值税之和，平均资产总额为期初和期末之和的算术平均值。

(2) 资本保值增值率。该指标反映企业净资产的变动状况，是企业发展能力的集中体现。

$$\text{资本保值增值率}=\frac{\text{期末所有者权益(净资产)}}{\text{期初所有者权益(净资产)}}\times 100\%$$

(3) 资产负债率。该指标既反映企业经营风险的大小，又反映企业利用债权人提供的资金从事经营活动的能力。

$$\text{资产负债率}=\frac{\text{期末负债总额}}{\text{期末资产总额}}\times 100\%$$

(4) 流动资产周转率。该指标是指一定时期内流动资产完成的周转次数，反映工业企业投入的流动资金周转速度。

$$\text{流动资产周转率}=\frac{\text{销售收入}}{\text{流动资产平均余额}}$$

其中：流动资产平均余额为期初与期末流动资产之和的算术平均值。

(5) 成本费用利润率。该指标反映工业生产成本及费用投入的经济效益，同时也反映企业降低成本的经济效益。

$$\text{成本费用利润率}=\frac{\text{利润总额}}{\text{成本费用总额}}\times 100\%$$

其中：成本费用总额为产品销售成本、销售费用、管理费用、财务费用之和。

(6) 全员劳动生产率。该指标反映企业的生产效率和劳动投入的经济效益。

$$\text{全员劳动生产率(元/人)}=\frac{\text{工业增加值}}{\text{全部职工平均人数}}$$

由于工业增加值是按现行价格计算的，而职工人数不含价格因素，因此应将增加值价格因素予以消除。可采用工业总产值价格指数近似地加以调整。具体做法是：用报告期现价工业总产值除以报告期不变价工业总产值得出报告期价格比系数。用报告期价格比系数再除以基期价格比系数，就得出工业总产值价格指数。如果报告期或基期更换不变价格，则应连续计算每年的价格指数，然后连乘求得。之后计算报告期可比价全员劳动生产率。

(7) 产品销售率。该指标反映工业产品生产已实现销售的程度，是分析工业产销衔接情况、研究工业产品满足社会需求程度的指标。

$$\text{产品销售率}=\frac{\text{工业销售产值}}{\text{工业总产值}}\times 100\%$$

2. 工业经济效益综合指数

工业经济效益综合指数是综合衡量工业经济效益各个方面在数量上总体水平的一种特殊相对数，是反映工业经济运行质量的综合指标。它是以各单项工业经济效益指标报告期实际数值分别除以该项指标的全国标准值并乘以各自权数，加总后除以总权数求得。

其计算公式为

$$\text{工业经济效益综合指数}=\sum\left(\frac{\text{某项经济效益指标报告期数值}}{\text{该项指标全国标准值}}\times\text{权数}\right)\div\text{总权数}$$

权数是根据七项考核指标在综合经济效益中的重要程度，用专家调查法确定的。其中：总资产贡献率20，资本保值增值率16、资产负债率12、流动资产周转率15、成本费用利润率14、全员劳动生产率10、产品销售率13。

各项指标的标准值，是参考我国近期工业经济指标的实际水平及一般标准确定的。

（三）财政部颁布实施的企业经济效益评价指标体系

（1）销售利润率。销售利润率反映企业销售收入的获利水平。计算公式为

$$\text{销售利润率}=\frac{\text{利润总额}}{\text{产品销售净收入}}\times100\%$$

产品销售净收入：指扣除销售折让、销售折扣和销售退回之后的销售净额。

（2）总资产报酬率。总资产报酬率用于衡量企业运用全部资产获利的能力。计算公式为

$$\text{总资产报酬率}=\frac{(\text{利润总额}+\text{利息支出})}{\text{平均资产总额}}\times100\%$$

$$\text{平均资产总额}=\frac{(\text{期初资产总额}+\text{期末资产总额})}{2}$$

（3）资本收益率。资本收益率是指企业运用投资者投入资本获得收益的能力。计算公式为

$$\text{资本收益率}=\frac{\text{净利润}}{\text{实收资本}}\times100\%$$

（4）资本保值增值率。资本保值增值率主要反映投资者投入企业的资本完整性和保全性。计算公式为

$$\text{资本保值增值率}=\frac{\text{期末所有者权益总额}}{\text{期初所有者权益总额}}\times100\%$$

资本保值增值率＝100%，为资本保值；资本保值增值率大于100%，为资本增值。

（5）资产负债率。资产负债率用于衡量企业负债水平高低情况。计算公式为

$$\text{资产负债率}=\frac{\text{负债总额}}{\text{资产总额}}\times100\%$$

（6）流动比率。流动比率衡量企业在某一时点偿付即将到期债务的能力，又称短期偿债能力比率。计算公式为

$$\text{流动比率}=\frac{\text{流动资产}}{\text{流动负债}}\times100\%$$

速动比率：是指速动资产与流动负债的比率，它是衡量企业在某一时点上运用随时可变现资产偿付短期债务的能力。速动比率是对流动比率的补充。计算公式为

$$速动比率=\frac{速动资产}{流动负债}\times 100\%$$

$$速动资产=流动资产-存货$$

(7) 应收账款周转率。应收账款周转率也称收账比率,用于衡量企业应收账款周转快慢。计算公式为

$$应收账款周转率=\frac{赊销净额}{平均应收账款余额}\times 100\%$$

$$赊销净额=销售收入-现销收入-销售退回、折让、折扣$$

由于企业赊销资料作为商业机密不对外公布,所以,应收账款周转率一般用赊销和现销总额,即销售净收入。

$$平均应收账款余额=(期初应收账款余额+期末应收账款余额)\div 2$$

(8) 存货周转率。存货周转率用于衡量企业在一定时期内存货资产的周转次数,反映企业购、产、销平衡的效率的一种尺度。计算公式为

$$存货周转率=\frac{产品销售成本}{平均存货成本}\times 100\%$$

$$平均存货成本=(期初存货成本+期末存货成本)\div 2$$

(9) 社会贡献率。社会贡献率是衡量企业运用全部资产为国家或社会创造或支付价值的能力。计算公式为

$$社会贡献率=\frac{企业社会贡献总额}{平均资产总额}\times 100\%$$

企业社会贡献总额:企业为国家或社会创造或支付的价值总额,包括工资(含奖金、津贴等工资性收入)、劳保退休统筹及其他社会福利支出、利息支出净额、应交增值税、应交产品销售税金及附加、应交所得税及其他税收、净利润等。

(10) 社会积累率。社会积累率指衡量企业社会贡献总额中多少用于上缴国家财政。计算公式为

$$社会积累率=\frac{上缴国家财政总额}{企业社会贡献总额}\times 100\%$$

上缴国家财政总额:应交增值税、应交产品销售税金及附加、应交所得税及其他税收等。

(四) 企业绩效评价指标体系

中华人民共和国财政部、国家经济贸易委员会、中央企业工作委员会、中华人民共和国劳动和社会保障部、中华人民共和国国家发展和改革委员会于 2002 年对 1999 年印发的《国有资本金效绩评价操作细则》进行了修订,修改制定了《企业效绩评价操作细则(修订)》,以财统〔2002〕5 号通知印发,在国有企业执行。

根据《企业效绩评价操作细则(修订)》的规定,企业效绩评价指标体系由基本指标、修正指标、评议指标三个层次 28 个指标构成。

1. 基本指标

(1) 净资产收益率＝净利润/平均净资产。

(2) 总资产报酬率＝(利润总额＋利息支出)/平均资产总额。

(3) 总资产周转率＝主营业务收入净额/平均资产总额。

(4) 流动资产周转率＝主营业务收入净额/平均流动资产总额。

(5) 资产负债率＝负债总额/资产总额。

(6) 已获利息倍数＝(利润总额＋利息支出)/利息支出。

(7) 销售(营业)增长率＝本年主营业务收入增长额/上年主营业务收入总额。

(8) 资本积累率＝本年所有者权益增长额/年初所有者权益。

2. 修正指标

(1) 资本保值增值率＝扣除客观增减因素后的年末所有者权益/年初所有者权益。

(2) 主营业务利润率＝主营业务利润/主营业务收入净额。

(3) 盈余现金保障倍数＝经营现金净流量/净利润。

(4) 成本费用利润率＝利润总额/成本费用总额。

(5) 存货周转率＝主营业务成本/存货平均余额。

(6) 应收账款周转率＝主营业务收入净额/应收账款平均余额。

(7) 不良资产比率＝年末不良资产总额/年末资产总额。

(8) 现金流动负债比率＝年经营现金净流量/年末流动负债。

(9) 速动比率＝速动资产/流动负债。

(10) 三年资本平均增长率(所有者权益)。

(11) 三年销售平均增长率(主营业务收入)。

(12) 技术投入比率＝当年技术转让费支出与研发投入/当年主营业务收入净额。

3. 评议指标

(1) 经营者基本素质。

(2) 产品市场占有能力(服务满意度)。

(3) 基础管理水平。

(4) 发展创新能力。

(5) 经营发展战略。

(6) 在岗员工素质。

(7) 技术装备更新水平(服务硬环境)。

(8) 综合社会贡献。

(五) 能源经济效益统计指标

1. 能源经济效益

能源经济效益是指投入能源与产出的经济效果的比较。能源经济效益体现在能源系统流程的始终，它涉及生产领域，也涉及消费领域和流通领域，它存在于任何经济形态之中，可以从多方面、多层次了解能源系统流程各个领域中的投入与产出情况、能源合理有效利用程度。

2. 能源弹性系数

能源弹性系数亦称能源弹性。弹性可简单地理解为反直性或敏感性,它是衡量某一变量的变化所引起的另一相关变量的相对变化的指标。通常表示为在某一变量变化 1% 时,另一变量变化的相对程度。

能源与很多经济现象存在着相互依存和制约的数量关系,为研究能源在社会发展中的作用,分析能源商生产的增长、能源消费的增长对经济增长的影响,可分别计算能源生产弹性系数和能源消费弹性系数,通过这些指标可反映出能源的发展与社会经济的发展相互制约的关系以及发展趋势和规律。

能源弹性系数的种类及其计算如下。

(1) 能源生产弹性系数。能源生产弹性系数也称能源生产增长系数,是指能源生产增长率与国民生产增长率的比值,是指反映能源生产量的增长与国民经济增长之间关系的指标,计算公式为

$$\text{能源生产弹性系数}=\frac{\text{能源生产总量年平均增长速度}}{\text{国民经济年平均增长速度}}$$

国民经济指标,根据研究问题的不同需要,可采用国内生产总值、工业总产值等指标。与国外对比,多采用国内生产总值指标。能源生产总量是指一定时期内一个国家(地区)一次能源生产的总和,一次能源生产量包括原煤、原油、天然气、水电及其他动力能(如风能、地热能等)发电量,不包括低热值燃料生产量、生物质能、太阳能的利用和由一次能源加工转换而成的二次能源产量,因二次能源是由一次能源加工转换而来的,它只改变能量形态,并不增加能源的资源量。

(2) 能源消费弹性系数。能源消费弹性系数是指能源消费增长率与国民增长率的比值,是指反映能源消费总量与国民经济增长之间关系的指标,计算公式为

$$\text{能源消费弹性系数}=\frac{\text{能源消费总量年平均增长速度}}{\text{国民经济年平均增长速度}}$$

能源消费量和国民经济指标的选择,根据研究问题的不同需要,可采用能源消费总量和国内生产总值、工业部门能源消费量和工业总产值、工业增加值等指标。能源消费总量是指一定时期内一个国家(地区)物质生产部门和生活消费的各种能源的总和,它包括消费的原煤和原油及其制品、天然气、电力,不包括低热值燃料、生物质能、太阳能等的利用,可以分为终端能源消费量、能源加工转换损失和损失量三个组成部分。

第四节　常用统计分类标准

一、经济成分的划分

经济成分分为公有经济(包括国有经济和集体经济)和非公有经济(包括私有经济、港澳台经济、外商经济)。

(1) 国有经济单位,指生产资料归国家所有的各种企业、事业单位,以及各级国家机关、人民团体等单位。

(2) 集体经济单位,指生产资料归公民集体所有的各种企业、事业单位。包括农村各

种经济组织经营的农、林、牧、副、渔业，乡、村经营的企业、事业单位；城市、县、镇以及街道举办的集体经济性质的企业、事业单位。

(3) 私营经济单位，指生产资料归公民私人所有的单位。包括私营独资企业、私营合伙企业和私营有限责任公司。

(4) 联营经济单位，指不同所有制性质的企业之间或者企业、事业单位之间共同投资组成新的经济实体。包括紧密型联营企业、半紧密型联营企业和松散型联营企业。

(5) 股份制经济单位，指全部注册资本由全体股东共同出资，并以股份形式投资举办企业。主要包括股份有限公司和有限责任公司。

(6) 外商投资经济单位，指外国投资者根据中华人民共和国有关涉外经济的法律、法规，以合资、合作或独资的形式在中国境内开办企业。包括中外合资经营企业、中外合作经营企业和外资企业。

(7) 中国港澳台投资经济单位，指港、澳、台地区投资者参照中华人民共和国有关涉外经济的法律、法规，以合资、合作或独资的形式在内地举办企业。包括合资经营企业、合作经营企业和独资企业。

二、三次产业的划分

三次产业划分范围如下。

第一产业是指农、林、牧、渔业。

第二产业是指采矿业，制造业，电力、热力及水生产和供应业，建筑业。

第三产业是指除第一、二产业以外的其他行业。第三产业包括：批发和零售业，交通运输、仓储和邮政业，住宿和餐饮业，信息传输、软件和信息技术服务业，金融业，房地产业，租赁和商务服务业，科学研究和技术服务业，水利、环境和公共设施管理业，居民服务、修理和其他服务业，教育、卫生和社会工作，文化、体育和娱乐业，公共管理、社会保障和社会组织，国际组织。

三次产业分类如表 2-1 所示。

表 2-1　三次产业分类

三次产业分类类　别	《国民经济行业分类》(GB/T 4754—2011)类别名称及代码		
	门类	大类	类别名称
第一产业	A		农、林、牧、渔业
		01	农业
		02	林业
		03	畜牧业
		04	渔业
		05	农、林、牧、渔服务业

续表

三次产业分类类　别	《国民经济行业分类》(GB/T 4754—2011)类别名称及代码		
	门类	大类	类 别 名 称
第二产业	B		采矿业
		06	煤炭开采和洗选业
		07	石油和天然气开采业
		08	黑色金属矿采选业
		09	有色金属矿采选业
		10	非金属矿采选业
		11	开采辅助活动
		12	其他采矿业
	C		制造业
		13	农副食品加工业
		14	食品制造业
		15	酒、饮料和精制茶制造业
		16	烟草制品业
		17	纺织业
		18	纺织服装、服饰业
第二产业		19	皮革、毛皮、羽毛及其制品和制鞋业
		20	木材加工和木、竹、藤、棕、草制品业
		21	家具制造业
		22	造纸和纸制品业
		23	印刷和记录媒介复制业
		24	文教、工美、体育和娱乐用品制造业
		25	石油加工、炼焦和核燃料加工业
		26	化学原料和化学制品制造业
		27	医药制造业
		28	化学纤维制造业
		29	橡胶和塑料制品业
		30	非金属矿物制品业
		31	黑色金属冶炼和压延加工业
		32	有色金属冶炼和压延加工业
		33	金属制品业
		34	通用设备制造业
		35	专用设备制造业
		36	汽车制造业
		37	铁路、船舶、航空航天和其他运输设备制造业
		38	电气机械和器材制造业
		39	计算机通信和其他电子设备制造业
		40	仪器仪表制造业
		41	其他制造业
		42	废弃资源综合利用业
		43	金属制品、机械和设备修理业

续表

三次产业分类类　　别	《国民经济行业分类》(GB/T 4754—2011)类别名称及代码		
	门类	大类	类 别 名 称
第二产业	D		电力、热力及水生产和供应业
		44	电力、热力生产和供应业
		45	燃气生产和供应业
		46	水的生产和供应业
	E		建筑业
		47	房屋建筑业
		48	土木工程建筑业
		49	建筑安装业
		50	建筑装饰和其他建筑业
第三产业	F		批发和零售业
		51	批发业
		52	零售业
	G		交通运输、仓储和邮政业
		53	铁路运输业
		54	道路运输业
		55	水上运输业
		56	航空运输业
		57	管道运输业
		58	装卸搬运和运输代理业
		59	仓储业
		60	邮政业
	H		住宿和餐饮业
		61	住宿业
		62	餐饮业
	I		信息传输、软件和信息技术服务业
		63	电信、广播电视和卫星传输服务
		64	互联网和相关服务
		65	软件和信息技术服务业
	J		金融业
		66	货币金融服务
		67	资本市场服务
		68	保险业
		69	其他金融业
	K		房地产业
		70	房地产业
	L		租赁和商务服务业
		71	租赁业
		72	商务服务业

续表

三次产业分类类别	《国民经济行业分类》(GB/T 4754—2011)类别名称及代码		
	门类	大类	类别名称
第三产业	M		科学研究和技术服务业
		73	研究和试验发展
		74	专业技术服务业
		75	科技推广和应用服务业
	N		水利、环境和公共设施管理业
		76	水利管理业
		77	生态保护和环境治理业
		78	公共设施管理业
	O		居民服务、修理和其他服务业
		79	居民服务业
		80	机动车、电子产品和日用产品修理业
		81	其他服务业
	P		教育
		82	教育
	Q		卫生和社会工作
		83	卫生
		84	社会工作
	R		文化、体育和娱乐业
		85	新闻和出版业
		86	广播、电视、电影和影视录音制作业
		87	文化艺术业
		88	体育
		89	娱乐业
	S		公共管理、社会保障和社会组织
		90	中国共产党机关
		91	国家机构
		92	人民政协、民主党派
		93	社会保障
		94	群众团体、社会团体和其他成员组织
		95	基层群众自治组织
	T		国际组织
		96	国际组织

三、大中小微型企业的划分

统计上大中小微型企业划分办法

(1) 根据工业和信息化部、国家统计局、国家发展改革委、财政部《关于印发中小企业划型标准规定的通知》(工信部联企业〔2011〕300 号),结合统计工作的实际情况,特制定

本办法。

（2）本办法适用对象为在中华人民共和国境内依法设立的各种组织形式的法人企业或单位。个体工商户参照本办法进行划分。

（3）本办法适用范围包括：农、林、牧、渔业，采矿业，制造业，电力、热力、燃气及水生产和供应业，建筑业，批发和零售业，交通运输、仓储和邮政业，住宿和餐饮业，信息传输、软件和信息技术服务业，房地产业，租赁和商务服务业，科学研究和技术服务业，水利、环境和公共设施管理业，居民服务、修理和其他服务业，文化、体育和娱乐业 15 个行业门类以及社会工作行业大类。

（4）本办法按照行业门类、大类、中类和组合类别，依据从业人员、营业收入、资产总额等指标或替代指标，将我国的企业划分为大型、中型、小型、微型等四种类型。具体划分标准见附表。

（5）企业划分由政府综合统计部门根据统计年报每年确定一次，定报统计原则上不进行调整。

（6）本办法自印发之日起执行，国家统计局 2003 年印发的《统计上大中小型企业划分办法（暂行）》（国统字〔2003〕17 号）同时废止。

附表：

统计上大中小微型企业划分标准

行业名称	指标名称	计量单位	大型	中型	小型	微型
农、林、牧、渔业	营业收入(Y)	万元	$Y\geqslant 20\ 000$	$500\leqslant Y<20\ 000$	$50\leqslant Y<500$	$Y<50$
工业 *	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 1\ 000$ $Y\geqslant 40\ 000$	$300\leqslant X<1\ 000$ $2\ 000\leqslant Y<40\ 000$	$20\leqslant X<300$ $300\leqslant Y<2\ 000$	$X<20$ $Y<300$
建筑业	营业收入(Y) 资产总额(Z)	万元 万元	$Y\geqslant 80\ 000$ $Z\geqslant 80\ 000$	$6\ 000\leqslant Y<80\ 000$ $5\ 000\leqslant Z<80\ 000$	$300\leqslant Y<6\ 000$ $300\leqslant Z<5\ 000$	$Y<300$ $Z<300$
批发业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 200$ $Y\geqslant 40\ 000$	$20\leqslant X<200$ $5\ 000\leqslant Y<40\ 000$	$5\leqslant X<20$ $1\ 000\leqslant Y<5\ 000$	$X<5$ $Y<1\ 000$
零售业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 300$ $Y\geqslant 20\ 000$	$50\leqslant X<300$ $500\leqslant Y<20\ 000$	$10\leqslant X<50$ $100\leqslant Y<500$	$X<10$ $Y<100$
交通运输业 *	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 1\ 000$ $Y\geqslant 30\ 000$	$300\leqslant X<1\ 000$ $3\ 000\leqslant Y<30\ 000$	$20\leqslant X<300$ $200\leqslant Y<3\ 000$	$X<20$ $Y<200$
仓储业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 200$ $Y\geqslant 30\ 000$	$100\leqslant X<200$ $1\ 000\leqslant Y<30\ 000$	$20\leqslant X<100$ $100\leqslant Y<1\ 000$	$X<20$ $Y<100$
邮政业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 1\ 000$ $Y\geqslant 30\ 000$	$300\leqslant X<1\ 000$ $2\ 000\leqslant Y<30\ 000$	$20\leqslant X<300$ $100\leqslant Y<2\ 000$	$X<20$ $Y<100$

续表

行业名称	指标名称	计量单位	大型	中型	小型	微型
住宿业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 300$ $Y\geqslant 10\ 000$	$100\leqslant X<300$ $2\ 000\leqslant Y<10\ 000$	$10\leqslant X<100$ $100\leqslant Y<2\ 000$	$X<10$ $Y<100$
餐饮业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 300$ $Y\geqslant 10\ 000$	$100\leqslant X<300$ $2\ 000\leqslant Y<10\ 000$	$10\leqslant X<100$ $100\leqslant Y<2\ 000$	$X<10$ $Y<100$
信息传输业*	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 2\ 000$ $Y\geqslant 100\ 000$	$100\leqslant X<2\ 000$ $1\ 000\leqslant Y<100\ 000$	$10\leqslant X<100$ $100\leqslant Y<1\ 000$	$X<10$ $Y<100$
软件和信息技术服务业	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 300$ $Y\geqslant 10\ 000$	$100\leqslant X<300$ $1\ 000\leqslant Y<10\ 000$	$10\leqslant X<100$ $50\leqslant Y<1\ 000$	$X<10$ $Y<50$
房地产开发经营	营业收入(Y) 资产总额(Z)	万元 万元	$Y\geqslant 200\ 000$ $Z\geqslant 10\ 000$	$1\ 000\leqslant Y<200\ 000$ $5\ 000\leqslant Z<10\ 000$	$100\leqslant Y<1\ 000$ $2\ 000\leqslant Z<5\ 000$	$Y<100$ $Z<2\ 000$
物业管理	从业人员(X) 营业收入(Y)	人 万元	$X\geqslant 1\ 000$ $Y\geqslant 5\ 000$	$300\leqslant X<1\ 000$ $1\ 000\leqslant Y<5\ 000$	$100\leqslant X<300$ $500\leqslant Y<1\ 000$	$X<100$ $Y<500$
租赁和商务服务业	从业人员(X) 资产总额(Z)	人 万元	$X\geqslant 300$ $Z\geqslant 120\ 000$	$100\leqslant X<300$ $8\ 000\leqslant Z<120\ 000$	$10\leqslant X<100$ $100\leqslant Z<8\ 000$	$X<10$ $Z<100$
其他未列明行业*	从业人员(X)	人	$X\geqslant 300$	$100\leqslant X<300$	$10\leqslant X<100$	$X<10$

说明:(1) 大型、中型和小型企业须同时满足所列指标的下限,否则下划一档;微型企业只需满足所列指标中的一项即可。

(2) 附表中各行业的范围以《国民经济行业分类》(GB/T 4754—2011)为准。带*的项为行业组合类别,其中,工业包括采矿业,制造业,电力、热力、燃气及水生产和供应业;交通运输业包括道路运输业,水上运输业,航空运输业,管道运输业,装卸搬运和运输代理业,不包括铁路运输业;信息传输业包括电信、广播电视和卫星传输服务,互联网和相关服务;其他未列明行业包括科学研究和技术服务业,水利、环境和公共设施管理业,居民服务、修理和其他服务业,社会工作,文化、体育和娱乐业,以及房地产中介服务,其他房地产业等,不包括自有房地产经营活动。

(3) 企业划分指标以现行统计制度为准。

① 从业人员,是指期末从业人员数,没有期末从业人员数的,采用全年平均人员数代替。

② 营业收入,工业、建筑业、限额以上批发和零售业、限额以上住宿和餐饮业以及其他设置主营业务收入指标的行业,采用主营业务收入;限额以下批发与零售业企业采用商品销售额代替;限额以下住宿与餐饮业企业采用营业额代替;农、林、牧、渔业企业采用营业总收入代替;其他未设置主营业务收入的行业,采用营业收入指标。

③ 资产总额,采用资产总计代替。

四、城乡的划分

以国务院关于市镇建制的规定和我国的行政区划为基础，以民政部门确认的居民委员会和村民委员会为最小划分单元，将我国的地域划分为城镇和乡村。

城镇是指在我国市镇建制和行政区划的基础上，划定的区域。城镇包括城区和镇区。

城区是指在市辖区和不设区的市中，划定的区域。城区包括：街道办事处所辖的居民委员会地域；城市公共设施、居住设施等连接到的其他居民委员会地域和村民委员会地域。

镇区是指在城区以外的镇和其他区域中，划定的区域。镇区包括：镇所辖的居民委员会地域；镇的公共设施、居住设施等连接到的村民委员会地域；常住人口在 3 000 人以上独立的工矿区、开发区、科研单位、大专院校、农场、林场等特殊区域。

乡村是指本规定划定的城镇以外的其他区域。

五、高技术产业统计分类

表 2-2　高技术产业统计分类目录

行业代码	行业名称	行业代码	行业名称
253	核燃料加工	376	航空航天器制造
2665	信息化学品制造	3761	飞机制造及修理
27	医药制造业	3762	航天器制造
2710	化学药品原药制造	3769	其他飞行器制造
2720	化学药品制剂制造	40	通信设备、计算机及其他电子设备制造业
2730	中药饮片加工	401	通信设备制造
2740	中成药制造	4011	通信传输设备制造
2750	兽用药品制造	4012	通信交换设备制造
2760	生物、生化制品的制造	4013	通信终端设备制造
2770	卫生材料及医药用品制造	4014	移动通信及终端设备制造
368	医疗仪器设备及器械制造	4019	其他通信设备制造
3681	医疗诊断、监护及治疗设备制造	402	雷达及配套设备制造
3682	口腔科用设备及器具制造	403	广播电视设备制造
3683	实验室及医用消毒设备和器具的制造	4031	广播电视节目制作及发射设备制造
3684	医疗、外科及兽医用器械制造	4032	广播电视接收设备及器材制造
3685	机械治疗及病房护理设备制造	4039	应用电视设备及其他广播电视设备制造
3686	假肢、人工器官及植(介)入器械制造	404	电子计算机制造
3689	其他医疗设备及器械制造	4041	电子计算机整机制造

续表

行业代码	行 业 名 称	行业代码	行 业 名 称
4042	计算机网络设备制造	4115	试验机制造
4043	电子计算机外部设备制造	4119	供应用仪表及其他通用仪器制造
405	电子器件制造	412	专用仪器仪表制造
4051	电子真空器件制造	4121	环境监测专用仪器仪表制造
4052	半导体分立器件制造	4122	汽车及其他用计数仪表制造
4053	集成电路制造	4123	导航、气象及海洋专用仪器制造
4059	光电子器件及其他电子器件制造	4124	农、林、牧、渔专用仪器仪表制造
406	电子元件制造	4125	地质勘探和地震专用仪器制造
4061	电子元件及组件制造	4126	教学专用仪器制造
4062	印制电路板制造	4127	核子及核辐射测量仪器制造
407	家用视听设备制造	4128	电子测量仪器制造
4071	家用影视设备制造	4129	其他专用仪器制造
4072	家用音响设备制造	4141	光学仪器制造
409	其他电子设备制造	4154	复印和胶印设备制造
411	通用仪器仪表制造	4155	计算器及货币专用设备制造
4111	工业自动控制系统装置制造	4190	其他仪器仪表的制造及修理
4112	电工仪器仪表制造	621	公共软件服务
4113	绘图、计算及测量仪器制造	6211	基础软件服务
4114	实验分析仪器制造	6212	应用软件服务

思 考 题

1. 简述统计设计的含义、内容及其设计原则。
2. 简述统计指标、指标体系的含义和设计原则。
3. 简述经济成分分类、三次产业分类、大中小规模分类、城乡分类、高新及产业分类。
4. 简述工业总产值、增加值、销售产值、国内生产总值、劳动生产率、经济效益。
5. 对你熟悉的领域或事物设计一套指标体系。例如,大学生在学期间综合评价指标体系等。
6. 简述总量指标及其分类。
7. 简述相对指标及其计算。

练 习 题

1. 认真阅读下列统计分析报告,指出报告中使用的统计指标的类型。

2017 年,我市实现工业增加值 120 亿元,比上年增长 8.9%。其中,国有工业企业及年销售收入 500 万元以上的非国有工业企业实现工业增加值 79 亿元,比上年增长8.8%;国有及国有控股企业实现工业增加值 56 亿元,占 70.9%;三个城区的工业增加值之比为

1∶1.6∶2;职工总人数20万人,其中,女职工7万人,占35%;工资总额为14.4亿元,职工平均工资7 200元/人,比上一年提高20%。

2. 某总公司所属三个子公司,2017年产值计划完成情况如表2-3所示。

表2-3　某公司2017年产值计划完成情况

子公司	第一季度实际产值/万元	第二季度					第二季度为第一季度/%
		计划产值		实际产值		计划完成/%	
		产值	比重	产值	比重		
甲	105	110		116			
乙	125	135				100	
丙	162			195		95.1	
合计	392						

要求：计算并填上表中空格数值。

3. 某纺织公司,2017年棉纱产量如表2-4所示。要求计算标准(32支纱)产量。

表2-4　2017年某纺织公司棉纱混合产量

棉纱支数	产量/千克	千锭时产量/千克	折算系统	标准实物产量/千克
10	10 274	46.4		
18	45 670	36.9		
20	53 286	32.0		
32	23 540	16.9		
36	22 300	14.3		
40	33 765	12.6		
合计	188 835	—	—	

第三章

统计调查

【统计知识 ABC】

[A]人口统计指标

1. 人口数 指一定时点、一定地区范围内的有生命的个人的总和。

年度统计的年末人口数 指每年 12 月 31 日 24 时的人口数。年度统计的全国人口总数内未包括台湾省和港澳同胞以及海外华侨人数。

2. 市镇总人口和乡村总人口

其定义有两种口径：第一种口径(按行政建制)市人口：市管辖区域内的全部人口(含市辖镇，不含市辖区县)；镇人口：县辖镇的全部人口(不含市辖镇)；县人口：县辖乡人口。第二种口径(按常住人口划分)市人口：设区的市的区人口和不设区的市所辖的街道人口；镇人口：不设区的市所辖镇的居民委员会人口和县辖镇的居民委员会人口；县人口：除上述两种人口以外的全部人口。

1952—1980 年数据为第一种口径的数据，1982 年以后的数据为第二种口径的数据。

3. 出生率(又称粗出生率) 指在一定时期内(通常为一年)平均每千人所出生的人数的比率，一般用千分率表示。计算公式为

$$出生率=年出生人数/年平均人数\times 1\,000‰$$

式中，出生人数指活产婴儿，即胎儿脱离母体时(不管怀孕月数)，有过呼吸或其他生命现象。年平均人数指年初、年底人口数的平均数，也可用年中人口数代替。

4. 死亡率(又称粗死亡率) 指在一定时期内(通常为一年)一定地区的死亡人数与同期平均人数(或期中人数)之比，一般用千分率表示。计算公式为

$$死亡率=年死亡人数/年平均人数\times 1\,000‰$$

5. 人口自然增长率 指在一定时期内(通常为一年)人口自然增加数(出生人数减死亡人数)与该时期内平均人数(或期中人数)之比，一般用千分率表示。计算公式为

$$\begin{aligned}人口自然增长率&=(本年出生人数-本年死亡人数)/年平均人数\times 1\,000‰\\&=人口出生率-人口死亡率\end{aligned}$$

6. 在业人口(又称就业人口) 指 15 周岁及 15 周岁以上人口中从事一定的社会劳动并取得劳动报酬或经营收入的人口。

7. 不在业人口 指 15 周岁及 15 周岁以上人口中未从事社会劳动的人口，包括在校学生、料理家务、待升学、市镇待业、离退休、退职、丧失劳动能力等非在业人口。

8. 总负担系数　指被抚养人口(0～14 岁和 65 岁以上人口)与 15～64 岁人口的比例。计算公式为

总负担系数=被抚养人口/15～64 岁人口×100%

9. 负担老年系数　指老年人口(65 岁以上人口)与 15～64 岁人口的比例。计算公式为

负担老年系数=老年人口/15～64 岁人口×100%

10. 负担少年系数　指少年儿童与 15～64 岁人口的比例。计算公式为

负担少年系数=少年儿童人口/15～64 岁人口×100%

(资料来源:http://www.stats.gov.cn/tjzd/tjzbjs/t20020327_14300.html.)

[B]第六次人口普查数据公布

一、人口总量

这次人口普查登记的全国总人口为 1 339 724 852 人,与 2000 年第五次全国人口普查相比,10 年增加 7 390 万人,增长 5.84%,年平均增长 0.57%,比 1990—2000 年的年平均增长率 1.07%下降 0.5 个百分点。数据表明,10 年来我国人口增长处于低生育水平阶段。

二、家庭户规模

这次人口普查,31 个省、自治区、直辖市共有家庭户 40 152 万户,家庭户人口 124 461 万人,平均每个家庭户的人口为 3.10 人,比 2000 年人口普查的 3.44 人减少 0.34 人。家庭户规模继续缩小,主要是由于我国生育水平不断下降、迁移流动人口增加、年轻人婚后独立居住等因素的影响。

三、性别构成

这次人口普查,男性人口占 51.27%,女性人口占 48.73%,总人口性别比由 2000 年人口普查的 106.74 下降为 105.20(以女性人口为 100.00)。

四、年龄构成

这次人口普查,0～14 岁人口占 16.60%,比 2000 年人口普查下降 6.29 个百分点;60 岁及以上人口占 13.26%,比 2000 年人口普查上升 2.93 个百分点,其中 65 岁及以上人口占 8.87%,比 2000 年人口普查上升 1.91 个百分点。我国人口年龄结构的变化,说明随着我国经济社会快速发展,人民生活水平和医疗卫生保健事业的巨大改善,生育率持续保持较低水平,老龄化进程逐步加快。

五、民族构成

这次人口普查,汉族人口占 91.51%,比 2000 年人口普查的 91.59%下降 0.08 个百分点;少数民族人口占 8.49%,比 2000 年人口普查的 8.41%上升 0.08 个百分点。少数民族人口 10 年年均增长 0.67%,高于汉族 0.11 个百分点。

六、各种受教育程度人口

这次人口普查,与 2000 年人口普查相比,每 10 万人中具有大学文化程度的由 3 611 人上升为 8 930 人,具有高中文化程度的由 11 146 人上升为 14 032 人;具有初中文化程度的由 33 961 人上升为 38 788 人;具有小学文化程度的由 35 701 人下降为 26 779 人。

文盲率(15 岁及以上不识字的人口占总人口的比重)为 4.08%,比 2000 年人口普查

的 6.72%下降 2.64 个百分点。

各种受教育程度人口和文盲率的变化,反映了 10 年来我国普及九年制义务教育、大力发展高等教育以及扫除青壮年文盲等措施取得了积极成效。

七、城乡构成

这次人口普查,居住在城镇的人口为 66 557 万人,占总人口的 49.68%;居住在乡村的人口为 67 415 万人,占 50.32%。同 2000 年人口普查相比,城镇人口比重上升 13.46 个百分点。这表明 2000 年以来我国经济社会的快速发展极大地促进了城镇化水平的提高。

八、地区分布

这次人口普查,东部地区人口占 31 个省(区、市)常住人口的 37.98%,中部地区占 26.76%,西部地区占 27.04%,东北地区占 8.22%。

与 2000 年人口普查相比,东部地区的人口比重上升 2.41 个百分点,中部、西部、东北地区的比重都在下降,其中西部地区下降幅度最大,下降 1.11 个百分点;其次是中部地区,下降 1.08 个百分点;东北地区下降 0.22 个百分点。

按常住人口分,排在前五位的是广东省、山东省、河南省、四川省和江苏省。2000 年人口普查排在前五位的是河南省、山东省、广东省、四川省、江苏省。

九、人口的流动

这次人口普查,居住地与户口登记地所在的乡镇街道不一致且离开户口登记地半年以上的人口为 26 139 万人,其中市辖区内人户分离的人口为 3 996 万人,不包括市辖区内人户分离的人口为 22 143 万人。同 2000 年人口普查相比,居住地与户口登记地所在的乡镇街道不一致且离开户口登记地半年以上的人口增加 11 700 万人,增长 81.03%;其中不包括市辖区内人户分离的人口增加 10 036 万人,增长 82.89%。这主要是多年来我国农村劳动力加速转移和经济快速发展促进了流动人口大量增加。

人口普查现场登记结束后,在全国随机抽取了 402 个普查小区进行事后质量抽查,通过与现场登记结果比对,这次普查的漏登率为 0.12%。

(资料来源:http://www.stats.gov.cn/zgrkpc/dlc/yw/t20110428_402722384.html.)

[C]人口平均预期寿命

人口平均预期寿命,指同时出生的一批人若按照某一时期各个年龄死亡率水平度过一生平均能够存活的年数,是综合反映人们健康水平的基本指标。

2010 年第六次全国人口普查资料表明,随着我国社会经济的快速发展,人民生活水平的不断提高以及医疗卫生保障体系的逐步完善,我国人口平均预期寿命继续延长,婴儿死亡率进一步下降,国民整体健康水平有较大幅度的提高。

一、平均预期寿命比 2000 年提高 3.43 岁

根据 2010 年第六次全国人口普查详细汇总资料计算,我国人口平均预期寿命达到 74.83 岁(表 3-1),比 2000 年的 71.40 岁提高 3.43 岁。分性别看,男性为 72.38 岁,比 2000 年提高 2.75 岁;女性为 77.37 岁,比 2000 年提高 4.04 岁。男女平均预期寿命之差与 10 年前相比,由 3.70 岁扩大到 4.99 岁。表明,在我国人口平均预期寿命不断提高的过程中,女性提高速度快于男性,并且两者之差在进一步扩大。这与世界其他国家平均预

期寿命的变化规律是一致的。

表 3-1　平均预期寿命变化　　岁

年份	合计	男	女	男女之差
1981	67.77	66.28	69.27	−2.99
1990	68.55	66.84	70.47	−3.63
2000	71.40	69.63	73.33	−3.70
2010	74.83	72.38	77.37	−4.99

2010 年世界人口的平均预期寿命为 69.6 岁，其中高收入国家及地区为 79.8 岁，中等收入国家及地区为 69.1 岁。可见，我国人口平均预期寿命不仅明显高于中等收入国家及地区，也大大高于世界平均水平，但比高收入国家及地区平均水平低 5 岁左右。从提高幅度看，2000—2010 年我国人口平均预期寿命提高 3.43 岁，比世界平均提高 2.4 岁快 1 岁左右。一般说来，平均预期寿命越高，提高速度越慢。但随着医药技术的发展和改善，一些平均预期寿命已处于较高水平的国家同期提高的速度也比较快，比如韩国提高 4.9 岁、新加坡 3.6 岁、巴西 3.0 岁、越南 2.9 岁、英国 2.7 岁、法国 2.4 岁、澳大利亚 2.5 岁、德国 2.1 岁等。

二、婴儿死亡率继续呈下降趋势

人口平均预期寿命的提高，是各年龄死亡率水平下降综合作用的结果，而婴儿死亡率（指同时出生的一批婴儿未能存活到 1 岁的比例）的下降起着尤为重要的作用。

2010 年我国婴儿死亡率为 13.93‰（表 3-2），比 2000 年的 28.38‰ 下降 14.45 个千分点，平均每年下降 1.45 个千分点。而 1990—2000 年 10 年间，我国婴儿死亡率下降 4.51 个千分点，平均每年下降 0.45 个千分点。可见，随着我国经济的发展，人民生活水平和妇幼保健服务水平的提高，我国的婴儿死亡率不仅继续呈下降趋势，而且下降速度加快。

表 3-2　婴儿死亡率变化　　‰

年　份	合计	男	女	男女之差
1990	32.89	32.36	33.48	−1.12
2000	28.38	23.92	33.75	−9.83
2010	13.93	13.62	14.30	−0.68

（资料来源：http://www.stats.gov.cn/tjgb/rkpcgb/qgrkpcgb/t20120921_402838652.html.）

第一节　统计调查概述

一、统计调查的含义

（一）统计调查的概念

对于社会经济现象，采集统计资料的主要手段就是统计调查。统计调查是根据统计研究的目的和要求，采用科学的方法，有组织、有计划地针对调查对象采集统计资料的工

作过程。

统计调查的基本任务是根据统计指标和指标体系,通过具体的统计调查,取得反映社会经济现象及其内在联系的原始统计资料。

采集统计资料的工作一般有两种:一种是对原始资料的采集,即直接对调查单位的情况进行观察或调查而取得一手资料;另一种是对已经加工过的现成资料的采集,称为次级资料的采集或二手资料的采集。

1. 一手资料的采集

一手资料的采集来源主要有两个渠道:一是调查或观察,二是实验。调查是取得社会经济数据的重要手段,其中有统计部门进行的统计调查,也有其他部门或机构为特定目的而进行的调查,如市场调查等;实验是取得自然科学数据的主要手段。在本章中,着重介绍取得社会经济数据的主要采集方式和方法。

2. 二手资料的采集

二手资料的采集既可以从报纸、图书、杂志、统计年鉴、网络等渠道获得,也可以从调查公司或数据库公司购买。近年来,互联网已经成为数据来源的重要渠道,几乎所有的政府机构和大公司都有自己的网站并提供公共访问端口,访问者可以从中获得有用的资料。如中华人民共和国国家统计局网(http://www.stats.gov.cn)、山东统计信息网(http://www.stats-sd.gov.cn)、中国经济信息网(http://www.cei.gov.cn)、华通数据中心(http://www.data.acmr.com.cn)、中国决策信息网(http://www.china-policy.com)、中国统计网(http://www.8sta.com)等。

利用二手资料对使用者来说既经济又方便,但使用时应注意统计数据的含义、计算口径和计算方法,以避免误用或滥用。同时,在引用二手资料时,一定要注明数据的来源,一方面体现严谨性,另一方面体现尊重他人的劳动。

(二)统计调查的基本要求

统计调查是统计工作的开始阶段,是统计整理和统计分析的前提。统计调查在整个统计工作中,担负着提供基础资料的任务,是一切统计资料的来源。统计工作的各个环节是紧密衔接、相互依存的。一般情况下,统计研究是从统计调查开始,通过接触实际情况,占有原始资料,取得感性认识,再经过对资料的系统整理和综合分析,提高到理性认识。统计调查工作的质量如何,直接影响到整个统计工作成果的质量。因此,准确、及时、全面、系统地提供统计调查资料,是对统计调查的基本要求。

(1) 准确性。统计调查所收集的资料必须准确可靠,符合实际情况。统计工作能否顺利完成任务,在很大程度上取决于所收集的资料是否准确。如果调查所收集的资料不准确,情况失实,那么根据这样的资料进行整理和分析,必将得出错误的结论。

(2) 及时性。要按照调查方案所要求的呈报统计资料的时间,及时调查、及时上报,以满足各部门对统计资料的需求。

(3) 全面性。按照调查计划的规定,对要调查的单位和项目资料,毫无遗漏地进行收集,以反映被调查事物的全貌。

(4) 系统性。收集的统计资料要符合事物的逻辑,便于整理、汇总。

二、统计调查的种类

根据不同的角度，统计调查可以有不同的分类。

（一）按统计调查对象的范围，分为全面调查和非全面调查

（1）全面调查，是对构成被研究总体的所有单位无一例外地进行调查。例如，要了解全国原油产量，就要对全国所有油田的原油产量进行调查登记。全面调查包括普查和全面统计报表。

（2）非全面调查，是对被研究总体的一部分单位进行的调查。例如，为了解我国公有制企业经济效益状况及经营管理的新情况、新问题，只选择部分公有制企业进行调查。非全面调查主要包括抽样调查、重点调查和典型调查等几种调查方法。

（二）按统计调查的登记时间是否连续，分为连续调查和不连续调查

（1）连续调查，是指随着被研究对象的变化，进行连续不断的登记，也称作经常性调查。例如，钢产量调查、人口出生（死亡）人数调查、连续生产过程中的产品产量或原材料消耗量调查等。

（2）不连续调查，也称一次性调查，是指间隔一段相当长的时间才对事物的变化进行一次调查。不连续调查获得的资料说明现象在某一天或某一时刻的数量。例如，人口数量、固定资产原值、生产设备数量等。

一般来说，两次调查间隔时间不超过 1 年的属于连续调查；间隔时间超过 1 年的属于非连续调查，它可以是周期的，也可以是不定期的。

（三）按统计调查的组织方式，分为统计报表和专门调查

（1）统计报表，是国家统计系统和业务部门为了定期取得系统、全面的基本统计资料而采用的一种收集资料的方式，目的在于掌握经常变动的、对国民经济有重大意义的指标的统计资料。

（2）专门调查，是为了了解和研究某种具体情况和问题而专门组织的统计调查，包括普查、重点调查、典型调查、抽样调查等。

三、统计调查方案

所谓统计调查方案，是指根据统计调查的目的和任务，按照调查对象的特点，对统计调查工作各方面和各环节所做的全面部署和安排。

统计调查方案设计是保证统计调查顺利进行的前提，也是准确、及时、系统地取得调查资料的重要的组织保证。一份完整的统计调查方案，应包括以下基本内容。

（一）调查目的

统计调查目的是进行统计调查所要达到的目标和需要解决的主要矛盾与问题。它回答的是为什么调查，要解决什么样的问题，调查具有什么样的社会经济意义，等等。只有

在调查目的明确之后,才能确定向谁调查、调查什么及采用什么方法进行调查。调查目的的表述应简明扼要。我国 2000 年第五次人口普查的目的是这样表述的:“为准确地查清第四次全国人口普查以来我国人口在数量、地区分布、构成和素质方面的变化,为科学地制定国民经济和社会发展战略与规划,统一安排人民的物质文化生活,检查人口政策执行情况,提供可靠的资料。”

(二)调查对象

明确调查对象就是确定调查对象和调查单位,解决向谁调查和由谁来具体提供资料的问题。

调查对象,就是根据调查的目的和要求,需要调查的社会经济现象的全体。是统计总体在统计调查阶段的具体化。

第三次全国工业普查的对象为我国境内(除台湾地区)的全部工业企业和附营工业单位。

调查单位就是组成调查对象的每一个单位。调查单位是调查项目的具体承担者,是总体单位在统计调查阶段的具体化。例如,2000 年第五次全国人口普查的调查单位是“每一个具有中华人民共和国国籍并在中华人民共和国境内常住的人”。

在统计调查过程中,还需要确定报告单位并把调查单位同报告单位区别开来。报告单位也称填报单位,是具体负责向上报告调查内容、提交统计资料的单位,一般是在行政管理或经济管理中具有一定独立性的单位,如国家机关、企事业单位、单位内部的分支机构等。调查单位与报告单位有时一致,有时不一致。例如,在工业企业普查中,每个工业企业既是调查单位又是报告单位;在工业企业生产设备调查中,调查单位是工业企业的每台生产设备,而报告单位则是每个工业企业。调查单位可以是人、社会组织,也可以是物、时间、空间或行为,但报告单位只能是人或社会组织,物、时间、空间或行为不能成为报告单位。

(三)调查内容

调查内容就是要确定调查项目和拟定调查表。

调查项目就是统计调查所要了解的调查单位的内容,是反映总体单位特征的标志在统计调查阶段的具体化。

例如,第三次全国工业普查的调查项目包括工业企业基本情况,财务状况,劳动情况,工业产品生产、销售、库存总值,主要工业产品产、销、存,附营工业单位生产情况,工业企业原材料、能源消费与库存,能源加工转换,工业企业科技活动情况,主要工业技术经济指标,主要工业产品生产能力,主要工业生产设备及其新旧程度(已安装设备),主要工业生产设备技术状况(已安装设备),主要工业产品质量,主要工业产品销售收支情况和工业企业部分合资基本经营情况 16 个方面。

一般来说,调查单位特征是非常多的,确定调查项目应该坚持以下几个方面的原则。

(1) 兼顾需要与可能。调查项目越多,调查费用也越高,时效性就越差。在进行统计调查时,则应选择能够切实满足调查目的要求而又可能得到答案的项目。

(2) 统一性。对调查项目的表述必须明确、易懂、统一,不能模棱两可,避免由于歧义

而造成登记差错。

(3) 调查项目之间要保持衔接。调查项目之间要保持一定的逻辑和计算关系，以便在实际调查和资料审核时进行核对，提高调查资料的质量。

(4) 发展和稳定相结合。随着统计调查对象的变化和发展，在不同时间进行的同类调查中，对调查项目应该进行适当的调整、补充和完善；同时，出于对统计资料可比性以及研究现象发展趋势和变化规律的考虑，调查项目也应尽可能保持稳定。

(四) 调查时间

在统计调查中，调查时间包括三个方面的含义。

(1) 调查资料的所属时间。如果所调查的是时期现象，就要明确规定调查资料的起止时间；如果是时点现象，就要明确调查资料的标准时点。

(2) 调查资料的登记时间。调查资料的登记时间是指对调查单位进行调查并取得调查资料的时间。

(3) 调查工作期限。调查工作期限是指从调查工作开始时起到调查工作结束时止所经历的全部时间，包括调查工作的准备、资料收集及资料报送等整个调查工作所需要的时间。这里需要明确的时间安排有：①总体方案论证、设计；②抽样方案设计；③问卷设计、测试、修改和定稿；④调查员的挑选和培训；⑤调查实施；⑥数据的整理、录入和分析；⑦调查报告的撰写；⑧有关鉴定、发布会和出版。

例如，2010 年 11 月 1 日零时，为第六次全国人口普查登记的标准时间(调查资料所属时间)，普查登记时间为 2010 年 11 月 1 日开始到 11 月 10 日。从成立领导机构、宣传发动、人员培训、资料登记、上报、汇总直至最后发布人口普查数据的时间，则属于调查工作期限。

(五) 调查预算

一项调查经费是一定的，合理制订经费预算方案，是提高调查质量同时节省费用的必要环节。一项完整的调查可能涉及的费用有：①调查方案的策划费与设计费；②抽样设计费；③问卷设计费；④问卷印刷、装订费；⑤调查实施费；⑥数据编码、录入费；⑦数据统计分析费；⑧调查报告撰写费；⑨办公费；⑩其他费用。

(六) 调查实施

为了保证整个统计调查工作顺利进行，在调查方案中还应该有一个周密的组织实施方案。主要内容包括调查机构成立、调查人员挑选和培训、调查资料报送办法、调查前的宣传发动、调查文件的准备以及试点工作等。

第二节　统计调查的组织方式

实际中常用的统计调查组织方式主要有统计报表、普查、抽样调查、重点调查和典型调查等。

一、统计报表

统计报表是按照国家有关法规规定,自上而下统一布置、自下而上逐级填报的一种调查组织方式。这种调查组织方式在我国政府统计工作中,经过几十年的改进和完善,已形成了一套比较完备的统计报告制度,它要求以原始数据为基础,按照统一的表式、指标、报送时间和报送程序填报,已成为国家和地方政府部门获取统计数据的主要统计调查组织方式。

统计报表类型多样。统计报表按调查范围可分为全面报表和非全面报表;按报送时间可分为日报、月报、季报和年报等;按报送受体可分为国家、部门、地方统计报表。

统计报表的主要特点和优点如下。

(1) 可以在调查前将报表布置到基层填报单位,以便它们根据报表的要求,及时建立、健全各种原始记录,使统计报表建立在可靠的基础上,做到资料准确、报送及时。

(2) 各级领导部门可以通过统计报表经常了解经济和社会发展变化的情况。

(3) 统计报表是经常性调查,有利于完整地积累资料,用于进行历史比较及较系统地分析、研究经济和社会发展变化的规律性。

统计报表是采集统计资料的重要工具,但也有其局限性,主要表现在以下几方面。

(1) 花费人力、物力大。

(2) 在层层上报的过程中容易出现偏差,影响数字的准确性。

(3) 社会经济现象复杂多变,统计报表实际上并不能收集到全部资料。

因此,在市场经济条件下,为了准确取得统计资料,还要根据调查目的和调查对象的特点,采取其他的调查方式。

二、普查

普查是为某一特定目的而专门组织的一次性全面调查方式,如人口普查、工业普查、经济普查等。世界各国一般都定期进行各种普查,普查适用于特定目的、特定对象,旨在收集有关国情国力的基本统计数据,为国家制定有关政策或措施提供依据。它主要用于收集处于某一时点状态上的社会经济现象的数量。

普查作为一种特殊的调查组织方式有以下几个特点。

(1) 普查通常是一次性或周期性的。

(2) 需要规定统一的标准时点。

(3) 普查数据较准确,规范化程度比较高,可作为抽样调查和其他调查的依据。

(4) 适用范围较窄,只能调查一些最基本或特定的现象。

三、抽样调查

(一) 抽样调查的概念

抽样调查也称样本调查,是非全面调查中的一种重要方法,它是按一定程序从所研究对象的全体(总体)中抽取一部分(样本)进行调查或观察,获取数据,并以此对总体的一定

目标量(参数)做出推断(如估计)。为了能更好地理解抽样调查的概念,还需要介绍两对与抽样调查有关的概念。

1. 目标总体与抽样总体

目标总体也可简称为总体,是指所要研究对象的全体,它由研究对象中所有性质相同的个体组成,组成总体的各个个体称为总体单元或单位。例如,我们要研究山东省个体商业情况,目标总体就是山东省所有从事商业活动的个体经营单位。

抽样总体是指从中抽取样本的总体。按理,抽样总体与目标总体完全一致,但实践中两者不一致的情况时常发生,这就有可能造成抽样框误差。

2. 抽样框与抽样单元

抽样总体的具体表现是抽样框。通常,抽样框是一份包含所有抽样单元的名单或目录,给每一个抽样单元编上一个号,就可以按一定的随机化程序进行抽样。对抽样框的基本要求是,抽样框中应该具有抽样单元名称和地理位置的信息,以便调查人员能够找到被选中的单元。

抽样单元是构成抽样框的基本要素。在大规模抽样调查中,当总体包含个体数量非常庞大时,直接对个体抽样在操作上往往是不方便的。为使概率抽样能够实施,同时也为了具体抽样便利,通常将总体划分成互斥且又穷尽的若干部分,每个部分都称为一个抽样单元,每个抽样单元都是由若干个体组成的集合。如果抽样单元只由一个个体组成就称为最小抽样单元。抽样单元的划分应视具体情况而定,它可以是自然形成的,也可以是人为划定的。例如,在人口变动抽样调查中可将县(区)、乡(街道)或村(居)民委员会作为抽样单元;在农产量抽样调查中可将人为分割的等面积的地块作为抽样单元。

(二)抽样调查的地位及作用

我国现行的统计法明确规定:统计调查应当以周期性的普查为基础,以经常性抽样调查为主体,以必要的统计报表、重点调查、综合分析等为补充,收集、整理基本统计资料。抽样调查与普查的关系是一种相互补充、相辅相成的关系。

抽样调查的作用:①经济性强。当总体较大时,应用抽样调查可以节约大量人力、物力、财力。②时效性高。与普查相比,抽样调查工作量小,可以较快提供调查结果。对于时效性要求较高的调查项目,通常采用抽样调查。③适应面广,可以承担普查无法胜任的项目。有些事物或客观现象,无法采用普查只能选择抽样调查。比如测试一宗炸弹的杀伤半径,普查是不可能的,只能采用抽样调查,抽几枚试炸一下即可。④准确性高,有助于提高调查数据质量。对抽样调查过程仔细监督、检查、指导,使得抽样调查数据比普查数据质量更高,总体总误差更小。

(三)抽样调查方法

抽样调查中的一个基本问题就是样本的抽取方法。在抽样调查中总体目标量的估计量及其精度都与具体抽样方法有关。抽样调查分为随机抽样和非随机抽样两种,其中统计中比较常用的是随机抽样,本节主要介绍随机抽样。最基本的随机抽样方法有以下五种,实际调查所用的方法通常是这五种方法的组合。

1. 简单随机抽样

简单随机抽样,也称为单纯随机抽样。从包含 N 个抽样单元的总体中抽取容量为 n 的简单随机样本,可以是从总体中逐个不放回地抽取 n 次,每次都是在尚未入样的单元中等概率抽取的,也可以是从总体中一次取得全部 n 个单元,只要保证全部可能的样本每个被抽到的概率都相等即可。

简单随机抽样是其他抽样方法的基础,因为它在理论上最容易处理,并且当总体包含的抽样单元数 N 不太大时,实施并不困难。但是当 N 很大时,实施就很困难,主要是编制一个包含全部 N 个抽样单元的抽样框通常很不容易。另外当 N 很大时,所抽到的样本单元往往很分散,使调查极不方便。因此在大规模的抽样调查中很少单独采用简单随机抽样。

2. 分层随机抽样

将总体中的抽样单元按某种原则划分为若干个子总体,每个子总体称为层。在每个层内独立地进行抽样,这样的抽样称为分层抽样。如果在每层内均采用简单随机抽样,就称为分层随机抽样。在分层抽样中,先根据层样本对层的参数进行估计,然后再将这些层估计加权平均或求和作为总体均值或总值的估计。

分层随机抽样特别适用于既要对总体参数进行估计也要对子总体(层)参数进行估计的情形。分层抽样实施和组织都比较方便,样本单元分布比较均匀。当层内单元指标差异较小而层间单元指标差异较大时,采用分层抽样可以大大提高估计的精度。例如,在居民收入调查中,按收入分布情况将居民分为最高收入层、高收入层、中等偏上层、中等收入层、中等偏下层、低收入层、最低收入层实施分层抽样,其估计精度就会比简单随机抽样有显著提高。

3. 二阶及多阶抽样

为抽样方便,有时我们把总体分成两个级别的抽样单元:初级抽样单元和次级抽样单元,总体由若干初级单元组成,每个初级抽样单元由若干次级抽样单元组成,先按某种方法在由初级单元构成的一级抽样框中抽样;然后在中选的初级单元中由次级单元构成的二级抽样框中抽样,抽样过程分为两个阶段,这种抽样方法称为二阶抽样。例如,在企业职工收入调查中,把企业作为初级抽样单元,职工作为次级抽样单元,先对企业进行抽样;再在被抽中企业内对职工进行抽样;然后对被抽中的职工进行调查,这就是二阶抽样。如果总体可以划分成多个级别的抽样单元,每一级别的抽样单元由若干下一级别的抽样单元组成,相应地存在多个级别的抽样框,抽样时先在一级抽样框中对一级单元抽样,再在中选的一级单元中对二级单元抽样,以此类推,这种抽样方法称为多阶抽样。多阶抽样实施方便,而且不需要对每个高级别的抽样单元建立关于低级别抽样单元的抽样框,调查费用也比较低。例如,在省抽县、县抽乡、乡抽村、村抽户的农产量四阶抽样中,凡未被抽中的县、乡、村就不必编制关于乡、村、户的抽样框。多阶抽样的主要缺点是估计量的结构比较复杂,估计量方差的估计也很复杂。

在二阶抽样中,如果对初级单元不再进行随机抽样,让所有的初级单元都入样,而在初级单元中对次级单元进行随机抽样,这样的二阶抽样就是分层随机抽样,层即初级单元。

4. 整群抽样

在二阶抽样中如果把初级抽样单元称作由次级抽样单元组成的群，在抽中的群内不再对次级单元进行抽样而是进行普查，那么这种抽样方法就称为整群抽样。当总体包含的次级单元为数众多且又缺少必要的档案资料因而无法直接对次级单元编制抽样框，而由次级单元组成的群的抽样框是现成的或者很容易编制时，常常采用整群抽样。整群抽样的优点是只需具备群即初级抽样单元的抽样框即可，无须具备关于次级单元的抽样框。整群抽样的效率与群的划分密切相关，如果总体划分成群后，群内差异小而群间差异大则估计精度就比较低。因此群的划分原则应是尽量扩大群内差异，使每个群都有较好的代表性。由此可知，划分群的原则正好和分层的原则相反。

5. 系统抽样

若总体中的抽样单元都按一定顺序排列，在规定的范围内随机抽取一个单元作为初始单元，然后按照一套事先定好的规则确定其他样本单元，这种抽样方法称为系统抽样。与其他几种抽样方法不同的是，这里只有初始单元是随机抽取的，其他样本单元都随着初始单元的确定而确定。最简单的系统抽样是在取得一个初始单元后按相等的间隔抽取后继样本单元，这种系统抽样称为等距抽样。等距抽样的优点是实施简单，整个样本中只是初始单元需随机抽取，其余单元皆由此决定。另外等距抽样有时甚至不需要编制抽样框，只需给出总体抽样单元的一个排列即可。如果对总体抽样单元的排列规则有所了解并加以正确利用，那么等距抽样就能达到相当高的精度。等距抽样的主要缺点是估计量精度的估计比较困难。

以上对几种常用的抽样方法做了简单的介绍。在实际运用中会有许多变化，如在某些抽样方法中，抽取样本单元可采用不放回抽样，也可采用放回抽样；可采用等概率抽样，也可采用不等概率抽样。在具体设计抽样方案时还应考虑多种因素，以决定抽样方法的选择和组合。

四、重点调查

重点调查是这样一种调查组织方式，它只从全部总体单位中选择少数重点单位进行调查，这些重点单位尽管在全部总体单位中出现的频数极少，但其某一数量标志却在所要研究的数量标志值总量中占有很大的比重。例如，要了解全国的钢铁生产总量，只要对产量很大的少数几个钢铁企业，如鞍钢、宝钢、首钢等进行调查，就可对全国的钢铁生产总量有个大致的认识。这几个产量很大的企业，构成了这次全国钢产量调查的重点单位，因为它们的钢铁产量在全国的钢铁生产总量中占有很大比重。重点调查的特点如下。

(1) 重点调查实质上是范围比较小的全面调查，它的目的是反映现象总体的基本情况。

(2) 重点调查单位的选择着眼于所研究现象主要标志的比重，因而它的选择带有主观因素。

(3) 重点单位对于总体来说具有代表性，但不能拿来推断总体总量。

当调查任务只要求掌握基本情况，调查对象又具有明显的重点单位时，可采用重点调查的方式。

五、典型调查

典型调查是从全部总体单位中选择一个或几个有代表性的单位进行深入细致调查的一种调查组织方式。典型调查的目的是通过典型单位具体生动、形象的资料来描述或揭示事物的本质或规律,因此所选择的典型单位应能反映所研究问题的本质属性或特征。例如,要研究工业企业的经济效益问题,可以在同行业中选择一个或几个经济效益突出的单位做深入细致的调查,从中找出经济效益好的原因和经验。典型调查的特点如下。

(1) 调查单位是根据调查的目的任务,在对现象总体进行全面分析的基础上有意识地选择出来的,它的选择具有较强的主观性。

(2) 典型调查是一种比较灵活的调查方式,既可以用于收集原始资料,又可以用于了解、掌握各种生动具体的情况。

(3) 利用典型调查资料做数量上的推算无法估计其误差,推算结果只是一个近似值。

典型调查主要用于定性研究,调查结果一般不能推断总体。

第三节 统计调查的方法

不论采用哪种方式组织调查,都要运用具体的数据收集方法去采集统计数据。归纳起来,数据收集方法有询问调查和观察与实验两大类。

一、询问调查

询问调查是调查者与被调查者直接或间接接触以获得数据的一种方法。具体包括面访调查、邮寄调查、电话调查、计算机辅助调查、网络调查、小组座谈会、个别深度访问等。

(一) 面访调查

面访调查又称派员调查,是调查者与被调查者通过面对面交谈从而得到所需资料的调查方法,这又可分为标准式访问和非标准式访问两种。标准式访问又称结构式访问,是按照调查者事先设计好的,有固定格式的标准化问卷或表格,有顺序地依次提问,并由被调查做出回答。其优点是能够对调查过程加以控制,从而获得比较可靠的调查结果。非标准式访问又称非结构式访问,它事先不制作统一的问卷或表格,没有统一的提问顺序,调查者只是给一个题目或提纲,由调查者和被调查者自由交谈,从中获得所需资料。询问调查在市场和社会调查中常被采用。

(二) 邮寄调查

邮寄调查是通过邮寄、宣传媒体和专门场所等将调查表或问卷送至被调查者手中,由被调查者填写,然后将调查表寄回或投放到收集点的一种调查方法。这是一种标准化调查,其特点是,调查者和被调查者没有直接的语言交流,信息的传递完全依赖于调查表。邮寄调查在统计部门进行的统计报表及市场调查机构进行的问卷调查中经常使用。

（三）电话调查

电话调查是调查者利用电话同被调查者进行语言交流，从而获得信息的一种调查方法。该方法具有时效快、费用低等特点。随着电话的普及，电话调查也越来越广泛。电话调查可以按照事先设计好的问卷进行，也可以针对某一专门问题进行电话采访。电话调查所提问题要明确，且数量不宜过多。

（四）计算机辅助调查

计算机辅助调查调查也叫作计算机辅助电话调查，就是在电话调查时，调查的问卷、答案都由计算机显示，整个调查过程，包括电话拨号、调查记录、数据处理等也都借助于计算机来完成的一种调查方法。目前，计算机辅助调查已在一些发达国家和地区广泛应用，并已开发出了各种计算机辅助电话调查系统。

（五）网络调查

利用网络进行调查。网络调查有两种形式：一是利用互联网直接进行问卷方式收集一手资料；二是利用互联网的媒体功能，从互联网上收集二手资料。

（六）小组座谈会

小组座谈会也称集体访谈法，就是将一组被调查者集中在调查现场，让他们对调查的主题发表意见，从而获取资料的方法。参加座谈会的被调查者应是所调查问题的专家或有经验者，人数不宜太多，通常为6～10人，研究人员应对被调查者进行严格的甄别、筛选。讨论方式主要看主持人的习惯和爱好。这种方法能获取其他方法无法取得的资料，因为在彼此交流的环境里，被调查者相互影响、启发、补充，不断修正自己的观点，这就有利于研究者从中获得较为广泛深入的想法和意见，而且座谈会不会因为问卷过长而遭到拒访。

（七）个别深度访问

深度访问是一种一次只要一名被调查者参加的特殊的定性研究。“深访”暗示着要不断深入被调查者的思想中，努力发掘其行为的真实动机。深访是一种无结构的个人访问，调查者运用大量的追问技巧，尽可能让被调查者自由发挥，表达他的想法和感受。深度访问常用于动机研究，如消费者购买某种产品的动机等，以发掘被调查者非表面化的深层意见。这一方法最适用于研究隐私的问题，如个人隐私问题，或敏感问题，如政治性问题。对于那些不同人之间观点差异极大的问题，用小组讨论可能会把问题弄糟，这时也可采用深度访问法。

小组座谈会和个别深度访问属于定性方法，通常围绕一个特定的主题取得有关定性资料，此类方法和定量方法不同。定量方法是从总体中按随机方式抽取样本获得资料，其研究结果或结论可以进行推论。但定性研究着重于问题的性质和对未来趋势的把握，而不是对研究总体数量特征的推断。小组座谈会和个别深度访问主要用于市场调查和研究。

二、观察与实验

观察与实验是调查者通过直接的观察或实验获得数据的一种方法。

(一) 观察法

这是指就调查对象的行动和意识,调查者边观察边记录的收集信息的方法。这是一种可替代直接发问的方法。运用这种方法,训练有素的观察员或调查员到重要地点,利用感觉器官或设置一定的仪器,观测和记录人们的行为和举动。采用观察方法,由于调查被调查不是强行介入,被调查者无须任何反应,因而常常能在被观测者不察觉的情况下获得信息资料。

(二) 实验法

这是一种特殊的观察调查方法。实验法是在所设定的特殊实验场所、特殊状态下,对调查对象进行实验以取得所需资料的一种调查方法。根据场所不同,实验法可分为在室内进行的室内实验法和在市场或外部进行的市场实验法。室内实验法可用于广告认知的实验等,如在同日的同种报纸上,版面大小相同,分别刊登 A、B 两种广告,然后将其散发给读者,以测定其反应结果。市场实验法可用于消费者需求调查等,如企业让消费者免费使用一种新产品,以得到消费者对新产品看法的资料。

第四节 调 查 问 卷

一、问卷及设计标准

问卷(或调查表、统计表)是专门为从被调查者那里获得有关某个主题的信息而设计的一组或一系列问题。

一份设计优良的问卷应该是:

(1) 能有效地用来收集数据,同时尽可能减少误差和矛盾;

(2) 对被调查者有友好的界面(如果有访员参与,还应对访员有友好的界面);

(3) 应尽量减少由无回答所引起的追踪回访的数量,且便于数据编码和录入,从而尽量减少审核与插补的工作量,最终减少收集和处理数据所花费的费用和时间。

问卷中的问题应符合调查的信息需求,并为调查分析提供有用的信息。虽然必须满足所有的信息需求条件,但是每一个问题都必须有明确的被列入问卷的理由,还应该让大家知道问每个问题的原因,以及将如何使用收集到的信息。问题的措辞必须清楚。对被调查者来说,问题的排列顺序必须符合逻辑。必须对问题的提法进行精心设计,使被调查者容易理解,并能给出确切的回答。最后,在调查实施之前,还应使用认知法、焦点座谈、非正式测试等方法对问卷进行测试。

二、问卷的开发流程

问卷的开发流程如下。

（1）向数据用户和对象进行咨询。调查目标的陈述和信息需求的确定，便于清楚地了解数据用途，统计调查机构才能设计出符合使用者要求的问卷。

（2）参考以前相同主题的问卷。通过参考在相同或相似的其他调查中使用过的问题，可为草拟本次调查的问题提供良好的基础。

（3）起草问题。起草问题时必须考虑下列因素：①数据收集方法；②被调查者；③回答负担；④所收集数据的复杂性；⑤信息的机密性和敏感性；⑥调查结果与其他调查结果的可比性；⑦数据可靠程度；⑧其他因素，包括：数据的可用性、被调查者提供数据的意愿、无回答、访员、数据处理、管理要求，以及对问题类型、问题措辞、问题顺序、问卷格式以及回答指南等的实际要求。

（4）对问卷进行审议与修改。在对问卷进行测试前，对问卷进行内部审议。参加审议人员应包括调查主题设计专家、问卷设计专家、访员、被调查总体中的代表。

（5）对问卷进行测试与修改。测试旨在确定问题中存在的困难和错误；测试还能确定问题的顺序是否正确、问题的解释以及问题的指南是否清楚，被调查者对问卷的表面印象如何。用于问卷测试的方法如下。

① 认知法。认知访谈通常在“实验室”环境或在带有单面镜等监督设备的房间中进行，样本数量相对较少，少则12～15次，多则100次以上。认知检验的方法包括观察被调查者、边想边说式访谈、试探性问题、释义、信心评定。

② 焦点座谈。焦点座谈是由从研究总体中选出的若干对象对所选主题进行的非正式讨论。通过焦点座谈这种形式，使被调查对象或数据用户以及访员有机会把他们的观点纳入问卷设计过程。焦点座谈由一位通晓小组访谈技术和熟悉讨论题目的人来主持。一次座谈通常由6～12人参加，一般进行两个小时左右。焦点座谈通常以要求参与者阐述他们对问卷的总体反应开始；然后再讨论与问卷有关的特殊问题。

③ 非正式测试。非正式测试样本量可从20到100，访问应按与计划中的正式调查同样的方式进行。非正式测试与认知法和焦点座谈不同，它并不与被调查者进行讨论，被调查者只是完成问卷或访问。

④ 向被调查者了解情况。

⑤ 向访员了解情况。

⑥ 分裂样本测试。在确定两个或多个版本的问卷和问题哪个更好时，可进行分裂样本测试。最简单的分裂样本设计是将一半样本试用问卷的一个版本，而另一半样本试用问卷的另一个版本。

⑦ 试点调查。

（6）定稿。

三、问题措辞设计技巧

如果被调查者容易理解用词的含义，那么调查数据的质量就会相对较高。如果被调查者清楚地理解了所问的内容，他们将更愿意且能够提供正确的信息，而且有必要让被调查者按照问卷设计者的意图理解调查问题。

如果出现下列情况，问题的措辞就有可能歪曲调查结果，从而导致不正确的数据。

(1) 被调查者不理解问题中的措辞含义。

(2) 被调查者对措辞的解释不同于设计者的意图。

(3) 被调查者不熟悉问题中措辞所表达的概念。

为避免这些问题,下面介绍了在对问题进行措辞时必须遵循的一些基本准则。

(一) 措辞要简单

与被调查者清楚地进行沟通的最佳途径是使用简单的日常用词,要确保所使用的术语适合于被调查的总体。

在设计问题时,要始终考虑被调查者的语言能力,应尽量选择每个人都容易理解的词语。

一般来说,调查者应该尽量避免使用回答调查问题的人所不熟悉的技术性很强的专业术语或仅在本专业内使用的“行话”。如果需要使用这样的术语,应该向被调查者澄清或解释这些术语的定义。必须对新的或复杂的概念进行定义,以使所有的被调查者对问题有一致的理解。定义可以包含在相关问题之中,也可以将所有这类定义或解释集中在一起,作为被调查者指南列在问卷的适当位置,还可以另外列在一本批示性手册中。

(二) 定义缩略语或专业术语

对一般总体的调查,不应使用缩略语。对任何必须使用的缩略语或简称,都应该事先进行定义。

(三) 确保问题适用

尽力减轻被调查者的负担,确保只提与被调查者有关的问题。问卷设计者要确保被调查者有足够的知识来回答被提问的问题。

(四) 措辞要具体

措辞要尽量具体,以确保调查者能确切理解对他们的要求。一定要清楚地说明:

(1) 问题适用于哪些人;

(2) 问题所指的是哪一时间段;

(3) 回答中应包括或不应包括的信息;

(4) 答案应按哪种计量单位给出。

下面的问题乍一看可能认为非常简单而直接。

您的收入是多少?

然而再仔细一想,这个问题实际上并不那么容易回答。首先“您的”的含义不清。它究竟指的是被调查者的个人收入还是家庭收入?这必须搞清楚。其次,要求被调查者提供的是哪段时期的收入呢?上周、上个月还是去年?最后,被调查者应计算什么种类的收入?只是薪金和工资?还是薪金、工资并包括津贴与奖金?或许还包括其他来源的收入?

下面的两个问题提供了对问题提问的两种较好的方法。

去年，在交税和各种扣除之前，您全家各种来源的总收入是多少？

或

去年，在各种扣除之前，您全家总收入是多少？包括来自工资、薪金和所有其他来源的收入。

（五）避免意义双关的问题

如果在一项提问中包含了两项以上的内容，被调查者就很难回答。

（六）避免诱导性问题

问卷中提出的问题不能带有倾向性，而应保持中立。诱导性问题能误导调查回答并影响调查结果。

（七）避免使用双重否定

问卷中应避免使用包含双重否定的句子结构，因为被调查者可能不知道他们是应该回答同意还是回答不同意。例如，你赞不赞成政府不允许便利店出售酒的规定？

（八）减少敏感问题或隐私问题的影响

涉及敏感的以及被调查者隐私或带威胁性的问题可能引起“社会意愿倾向”性的偏差。被调查者对这类问题在选择答案时有这样一种趋势：他们宁愿选择更有益于他们的自尊，或跟社会理想准则更加一致的选项，也不愿把真实信仰表达出来，或者把他们的真实处境暴露出来。这可能会导致对想要测量的行为或活动的低报。

对被调查者而言，自填式问卷比访员协助式问卷受敏感问题的威胁小。同样，如果被调查者相信他们可以匿名回答的话，那么他们会更倾向于如实回答较为敏感的问题。

下面是几种以较小威胁的方法来提问敏感问题的技巧，能够增加被调查者提供诚实回答的可能性。

1.“每个人”法

可以用那些表明这种敏感行为并非不常见的这类词语来引入敏感问题。像许多人或大多数人这些术语都可用来引进敏感话题。

2.为回答辩护技巧

这种技巧用以认为某种行为是情有可原的之类的话为其辩解的办法来引入敏感问题。通过告诉被调查者某种行为可能有很多正当的理由，使他们在报告这种行为时感到轻松一些。

3.“即使一次”技术

这种方法的目的是在询问有关当前行为的情况前，先确定被调查者是否曾经做过某事，即使只做过一次。在提出若干前导问题之后，就可以询问被调查者与所关心的主题有关的当前行为的情况了。

4. 系列问题或热身法

先提几个问题来减轻被调查者对敏感话题的敏感程度,让他们做好回答的准备,真正关心的问题在几个热身问题之后出现。这种方法也可以把前面提到的几种技巧结合起来。

5. 设立档次

对于某些私人信息,如年龄、收入或有关不良行为频率等方面的敏感信息,如果对不同的答案划分几个档次,让被调查者选择一个档次而不是确切数字的话,那么他们可能更愿意回答。

(九) 确保问题读来顺口

调查的问题应尽可能简明,并使用调查总体易懂的日常用语。应以友善的第二人称(你或您)来称呼被调查者,并遵守基本的语法规则。

重要的检验方法是,当将问题大声读出来的时候听起来感觉怎样。问题应该听起来自然,像谈话一样,且容易被对方所理解。下面是一个没有遵循这种原则的例子。

您如何对信息提供转变过程中的心理学和社会学方面的作用的程度进行评定,比如说使用计算机化的交互式职业咨询系统,该系统由部门的区域性办公室向面临退休的职工提供的,是由区域性人才办公室认可的。

这个问题太长,用了复杂的语言和结构,听起来生硬、官僚,又不通顺——既难懂,又难答。

四、问题答项设计技巧

回答项目是针对提问项目所设计的答案。由于问卷中的问题有不同类型,所设计的答案类型和对被调查者的回答要求也是不同的。

问卷中的问题类型有两类:一类是开放式问题;另一类是封闭式问题。

(一) 开放式问题

开放式问题是指对问题的回答未提供任何具体的答案,由被调查者根据自己的想法自由做出回答,属于自由回答型。

开放式问题的优点是比较灵活,适合于收集更深层次的信息,特别适合于那些尚未弄清各种可能答案或潜在答案类型较多的问题。而且可以使被调查者充分表达自己的意见和想法,有利于被调查者发挥自己的创造性。其缺点是,由于会出现各种各样的答案,给调查后的资料整理带来一定困难。

(二) 封闭式问题

封闭式问题是指对问题事先设计出了各种可能的答案,供被调查者进行选择。每种答案称为一个选项,要求被调查者在这些选项中选择一个或几个作为回答。

封闭式问题的优点是:①由于被调查者只是简单地选择合适的选项,而不需要用自己的语言来陈述答案,这样他们就能很快、更容易地回答问题;②由于答案标准化,所以封

闭式问题收集起来的数据更容易整理和分析；③与开放式问题相比，封闭式问题的编码和录入也更加容易，费用也更省；④如果一个问题被用于多项调查中，用相同的回答选项有利于对被调查结果进行比较。

封闭式问题的缺点是：①对于有些问题要得到可能回答选项，要花费很多精力。②如果选项阐述不清楚，被调查者遇到的问题可能比开放式问题更多，即一旦设计有缺陷，被调查者就可能无法回答问题，从而影响调查质量。③由于封闭式问题的答案是选择回答型的，所以设计出的答案一定要穷尽和互斥，当有些问题答案不能穷尽时，可以加上“其他”一项。④由于封闭式问题事先列出了选项，而被调查者可能感觉是被迫从中选择答案，因为不管他们是否对所需回答的问题有自己的观点或相应的知识，他们都必须对问题给予回答。为避免这种情况，有时需要加进一个“不知道”或“不适用”的选项。⑤另一个潜在的问题是选项可能将所研究的问题过于简单化，使得被调查者不能对某个回答进行详细阐述。对于被调查者来说，他们只能从列在问题后的现象中进行选择。

封闭式问题有多种，最常用的是两项选择、多项选择、顺序选择、等级评定、双向列联表等形式。

1. 两项选择

两项选择问题是封闭式问题中最简单的一种。通常它是一个判断“是”或“否”的问题，用于将被调查者拆分成两个对立的组。

当某些问题不是对所有被调查者都适用时，两项选择问题作为筛选（或甄别）的问题是很有用的。在选项之后，经常直接出现“跳至×问题”的提示，这样，被调查者就不会被问及不适用于他的问题。

2. 多项选择

若提供给被调查者选择的选项多于两个即是多项选择问题，如果假定被调查者只能从这些选项中选择一个，则称为单选问题；如果允许被调查者可选择一个或多个答案（可以限制数量也可以不限制数量），则称为多选问题。对被调查者来说，一个问题究竟是单选还是多选，问卷中应该有提示。注意，问卷为保证选项的穷尽性也常列出“其他（请注明）”这样的选项。

3. 顺序选择

顺序选择问题要求被调查者在回答时，对所选的答案按要求的顺序或重要程度加以排列（可以限制数量也可以不限制数量）。

被调查者经常发现回答排序问题很困难，特别是在需排列的项目实际上很不同的情况下尤其如此。排序问题的一个缺点是不知道相邻次序之间差距有多大，而且差距也不可能都相等。因此，1 和 2 之间的差距不能假定为与 2 和 3 之间的差距相同。例如，如果列出 3 个项目进行排序，被调查者将把它们排序为 1、2 和 3。然而，也许排在第 1 和第 2 项之间非常接近，而与最后一个项的差距很大。这种信息不能通过一个简单的排序就能弄清。另一个问题是有时被调查者可能将两项或更多项排在同一位置上。排序问题还有一个缺点是被调查者可能无法将清单上的所有选项进行排序，被调查者仅选择部分选取项进行排序也是可能的。

4. 等级评定

对于等级评定问题,一般是要求被调查者对选项进行评级。在构造等级评定问题时,有几个必须考虑的问题。

首先要考虑的是,应该有多少选项? 选项可以少至 2 个,如只要求被调查者对某个陈述回答同意或不同意的情形。选项也可以多至 10 个,如在选项中每一个都按照 1(一点不重要)到 10(特别重要)来划分等级的情形。

其次要考虑的是是否需要包含中性选项的问题,如——既不是满意也不是不满意。在等级评定问题中,被调查者有选择中性选取项的倾向。如果不给被调查者提供一个选择中性选项的机会,他们就会被迫做出某一选择。

在构造等级评定问题时还有一个要考虑的问题是,是否应列入“不知道”或“不适用”这类选项。是否包括此类选项取决于所提出的问题。例如,在问到被调查者从来没有使用过的某种特别服务时,有必要包含一个“不适用”的选项。

5. 双向列联表

双向列联表就是将两类不同问题综合到一起,通过表格来表现。表的横向是一类问题,表的纵向是另一类问题。这种问题结构可以反映两方面因素的综合作用,提供单一类型问题无法提供的信息,同时也可以节省问卷的篇幅。

五、常见问卷格式

不同的调查问卷在具体结构、题型、措辞、版式等设计上会有所不同,但在结构上一般都由开头部分、甄别部分、主体部分和背景部分组成。

(一) 开头部分

开头部分一般包括问候语、填表说明和问卷编号等内容。

(1) 问候语。在自填式问卷中,写好问候语十分重要,它可以引起被调查者的重视,消除顾虑,激发参与意识,以争取他们的积极合作。

问候语要语气亲切、诚恳礼貌,文字要简洁准确,并在结尾处对被调查者的参与和合作表示感谢。

(2) 填表说明。在自填式问卷中要有详细的填表说明,让被调查者知道如何填写问卷,如何将问卷返回到调查者手中。这部分内容可以集中放在问卷的前面,也可以分散到各有关问题的前面。

(3) 问卷编号。问卷编号主要用于识别问卷以便校对检查、更正错误。

(二) 甄别部分

甄别也称为过滤,它是先对被调查者进行过滤,筛选掉不需要的部分;然后针对特定的被调查者进行调查。

通过甄别或过滤,一方面可以筛选掉与调查事项有直接关系的人,以达到避嫌的目的;另一方面,也可以确定哪些人是合格的被调查者,哪些人不是。甄别的目的是确保被调查者合格,能够作为该调查项目的代表,从而符合调查研究的需要。

（三）主体部分

主体部分是调查问卷的核心内容，包括所要调查的全部问题，主要由问题和答案所组成。

（四）背景部分

背景部分通常放在问卷的最后，主要是有关被调查者的一些背景资料。该部分所包含的各项问题，可使研究者根据背景资料对被调查者进行分类比较分析。

六、Excel在问卷调查中的应用

用Excel做调查问卷比简单的Word文档的调查表要好很多，如果直接用计算机收集数据的话，利用Excel做调查问卷，被调查对象的选择结果会被依次复制到Excel的另外一张工作表中保存起来，问卷后期的录入工作会简单很多，而且还可以利用函数和公式对收集到的数据进行后续的统计处理和分析，大大提高问卷收集和整理的速度与准确度。另外，在调查问卷界面的美观上，Excel也有独特的优势。

下面我们举例说明利用Excel制作调查问卷和存储调查数据的思路与方法。

【例3-1】 某市场调查公司要通过组织一次问卷调查，以获得目前济南市手机市场的需求及使用状况。本次调查利用Excel做调查问卷，调查人员用计算机直接收集数据，要求可以在共享下多人同时对其进行操作，所有人的结果都会分别保存，不会被覆盖。问卷的最终效果如图3-1所示，问卷调查结果的最终效果如图3-2所示。



图3-1　Excel调查问卷的最终效果

	A	B	C	D	E	F	G	H	I	J
1	调查结果数据表									
2	参加调查人数:									
3					使用过的手机品牌					
4	序号	性别	年龄	学历	诺基亚	三星	摩托罗拉	联想	其他	现有手机使用时间
5	1	2	3	4		TRUE	TRUE	FALSE	TRUE	3
6	2	1	2	1		TRUE	FALSE	FALSE	FALSE	3
7	3	2	2	1		TRUE	TRUE	TRUE	TRUE	3
8	4	2	2	1		TRUE	TRUE		TRUE	3
9	5	2	6	1		TRUE				3
10	6	2	6	1		TRUE				3
11	7	2	6	1		TRUE				0
12	8	2	6	1		TRUE				0
13										

图 3-2 问卷调查结果的最终效果

Excel 2013 中的操作步骤如下。

1. 录入问卷的开头部分

(1) 新建工作簿。

启动 Excel 2013,创建一个新的工作簿。

(2) 重命名工作表。

双击 Sheet1 工作表标签进入重命名状态,输入“问卷调查表”,然后按 Enter 键确认。

(3) 输入问卷开头部分。

在 A1 单元格内输入问卷标题,在 A2 单元格内输入问候语,在 A3 单元格内输入填写说明。

2. 录入问卷的主体部分

(1) 录入第一部分标题。

在 A4 单元格内录入调查的第一部分标题。

(2) 利用“窗体控件”建立性别、年龄、学历的调查问题。

① 在 Excel 2013 中调出窗体控件。

单击“文件”>“选项”,在弹出的“Excel 选项”窗口中(图 3-3)选中“快速访问工具栏”选项,在右侧的“从下列位置选择命令”下拉列表中选择“不在功能区的命令”,在下方出现的命令菜单中,选择“按钮(窗体控件)”选项,然后单击“添加”按钮,把其添加到快速访问工具栏,然后把“分组框(窗体控件)”“复选框(窗体控件)”“数值调节钮(窗体控件)”“选项按钮(窗体控件)”“组合框(窗体控件)”选项依次添加到快速访问工具栏。

最后单击“确定”按钮。这样就可以在左上角的快速访问工具栏里看到添加的窗体控件,如图 3-4 所示。

② 建立“性别”调查问题。单击快速访问工具栏上的“分组框”按钮,在 A5:H7 的位置拖动合适大小;选中分组框左上角的名字,把其修改为“性别”;再单击快速访问工具栏上的“选项按钮”,在“性别”分组框的左侧拖动到适当大小,选中其名字修改为“男”;复制该选项按钮,放在“性别”分组框的右侧,将其改名为“女”。

③ 建立“年龄”调查问题。可参照②的操作方法。

④ 建立“学历”调查问题。

单击快速访问工具栏上的“分组框”按钮,在 A14:H17 的位置拖动合适大小;选中分组框左上角的名字,把其修改为“学历”;再单击快速访问工具栏上的“组合框”按钮,将

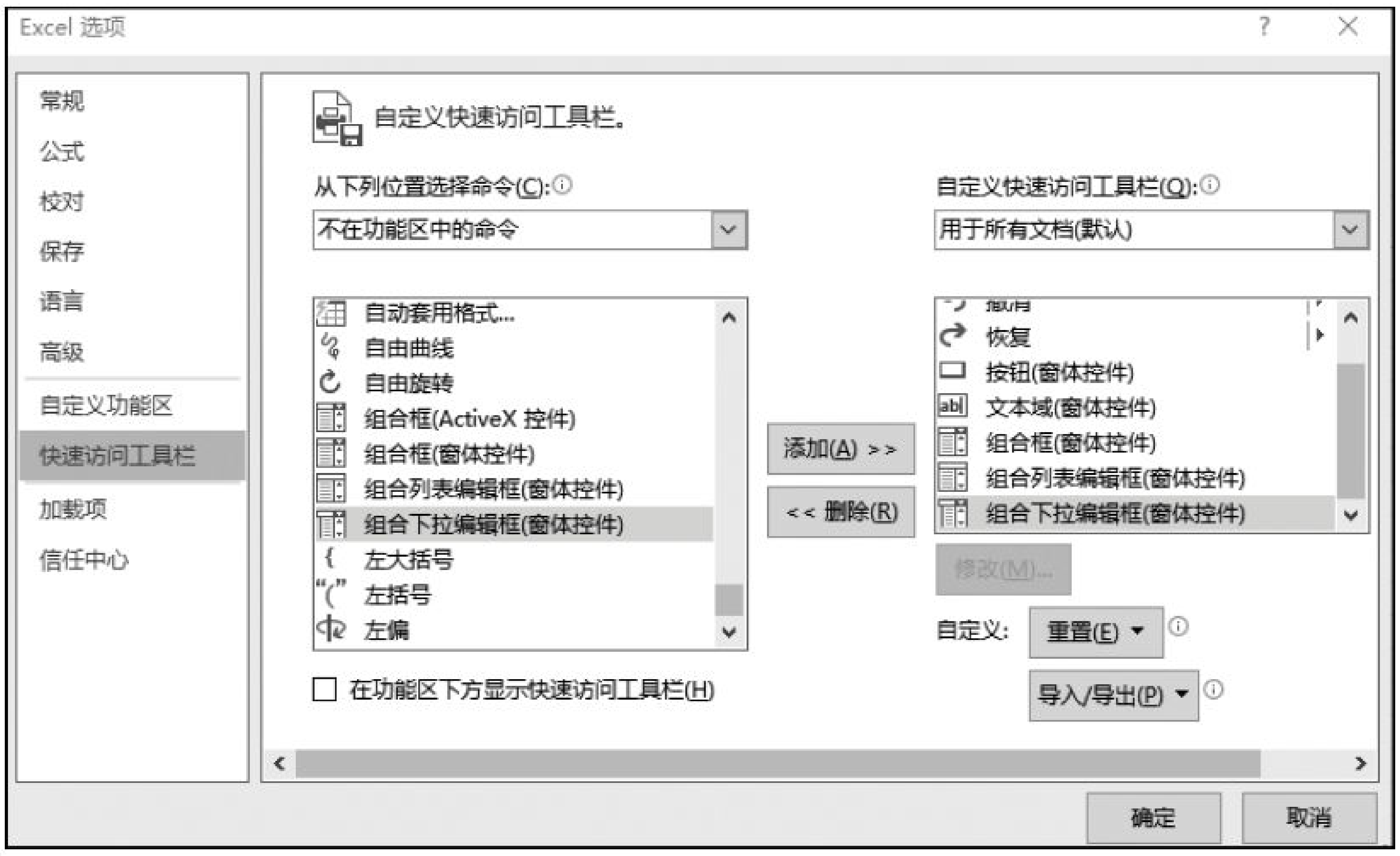


图 3-3　Excel 选项窗口



图 3-4　Excel 快速访问工具栏

其在“学历”分组框上拖动到合适大小。

（3）录入第二部分标题。

在 A19 单元格中录入调查的第二部分标题。

（4）利用“窗体控件”建立“使用手机品牌”“现有手机的已使用时间”的调查问题。

① 单击快速访问工具栏上的“分组框”按钮，在 A20：H26 的位置拖动合适大小；选中分组框左上角的名字，把其修改为“您使用过的手机品牌”；再单击快速访问工具栏上的“复选框”按钮，将其放在“您使用过的手机品牌”分组框的左侧，选中其名字将其重命名为“诺基亚”；按照上述方法依次建立“三星”“摩托罗拉”“联想”“其他”复选框。

② 单击快速访问工具栏上的“分组框”按钮，在 A27：H30 的位置拖动合适大小；选中分组框左上角的名字，把其修改为“现有手机的已使用时间”；再单击快速访问工具栏上的“数值调节钮”按钮，将其在“现有手机的已使用时间”分组框上拖动到合适大小。

③ 在 E29 单元格中输入“年”。

3．建立调查结果数据表

（1）单击 sheet2 工作表标签进入重命名状态，输入“调查结果表”，然后按“Enter”键确认。

（2）在调查结果表的 A1 单元格中录入标题“调查结果数据表”；在 A2 单元格中录入“参加调查人数”；在第 5 行和第 6 行对应单元格内分别录入调查结果数据的标题。

4．设置各窗体控件的格式

（1）设置性别单选按钮的格式。

右键单击“男”单选按钮，选择“设置控件格式”选项，弹出“设置控件格式”窗口，如

图 3-5 所示。



图 3-5 “设置控件格式”窗口

单击“控制”选项卡,将鼠标移到“单元格链接”文本框中,用鼠标选择“调查结果表”标签,再选中 B3 单元格。这样该组控件就跟调查结果表中的 B3 单元格建立的链接。单击“确定”按钮。

(2) 设置年龄单选按钮的格式。

右键单击“20 岁以下”单选按钮,选择“设置控件格式”选项,弹出“设置控件格式”窗口,如图 3-6 所示。



图 3-6 “设置控件格式”窗口

单击“控制”选项卡,将鼠标移到“单元格链接”文本框中,用鼠标选择“调查结果表”标签,再选中 C3 单元格。这样该组控件就跟调查结果表中的 C3 单元格建立的链接。单击“确定”按钮。

(3) 设置学历控件的格式。

首先建立学历控件的数据源:在“问卷调查表”的 K9:K17 区域依次输入图 3-7 所示

内容。

然后右键单击“学历”组合框控件，选择“设置控件格式”选项，弹出“设置控件格式”窗口，如图 3-8 所示。



图 3-7　学历控件数据源



图 3-8　“设置控件格式”窗口

单击“控制”选项卡，将鼠标移到“数据源区域”文本框中，用鼠标选择 K9：K17 区域；把鼠标移到“单元格链接”文本框内，单击“调查结果表”标签，选中 D3 单元格；在下拉显示项数中输入“5”。单击“确定”按钮。

（4）设置你使用过的手机品牌复选框的格式。

右键单击“诺基亚”复选框，选择“设置控件格式”选项，弹出“设置控件格式”窗口，如图 3-9 所示。



图 3-9　“设置控件格式”窗口

把鼠标移到“单元格链接”文本框内，单击“调查结果表”标签，选中 E3 单元格。

按照此方法,依次将“三星”“摩托罗拉”“联想”“其他”复选框分别与“调查结果表”的 F3、G3、H3、I3 单元格建立链接。

(5) 设置现有手机的已使用时间数值调节控件的格式。

右键单击数值调节钮控件,选择“设置控件格式”选项,弹出“设置控件格式”窗口,如图 3-10 所示。



图 3-10 “设置控件格式”窗口

将最小值设置为“1”,最大值设置为“10”,步长设置为“1”,单元格链接设置为本表的 D29 单元格;再单击“调查结果表”标签,在 J3 单元格中输入公式“=调查表!$D$29”,将其与数值调节控件建立链接。

5. 通过编辑“回答完成”按钮,实现调查数据的逐一自动保存

(1) 建立“回答完成”按钮。

单击快速访问工具栏上的“按钮”,在问卷主体部分的右下角,拖动到适当大小,选中其名称,将其重命名为“回答完成”。

(2) 实现调查数据的逐一自动保存。

一般来讲,实现调查数据的自动保存需要用 Excel 的 VBA 编程,但这对于没有学习过程序设计的非计算机人员来讲,存在很大的困难,下面我们介绍一种不需要 VBA 编程的方法。

右键单击“回答完成”按钮,选择“指定宏”选项,弹出“指定宏”窗口,如图 3-11 所示。

单击“编辑”按钮,进入“VB 编辑器”窗口,如图 3-12 所示。

此时,开始录制宏,单击工作簿的“调查结果数据表”标签,单击 A7 单元格,再依次单击 Excel“开始”菜单→插入→插入新工作表行;选择 B3:J3 单元格区域,右键单击选择“复制”选项;用鼠标选中 B7 单元格,右键选择“粘贴”选项,再单击“问卷调查表”标签,最后单击“保存”按钮。



图 3-11 “指定宏”窗口

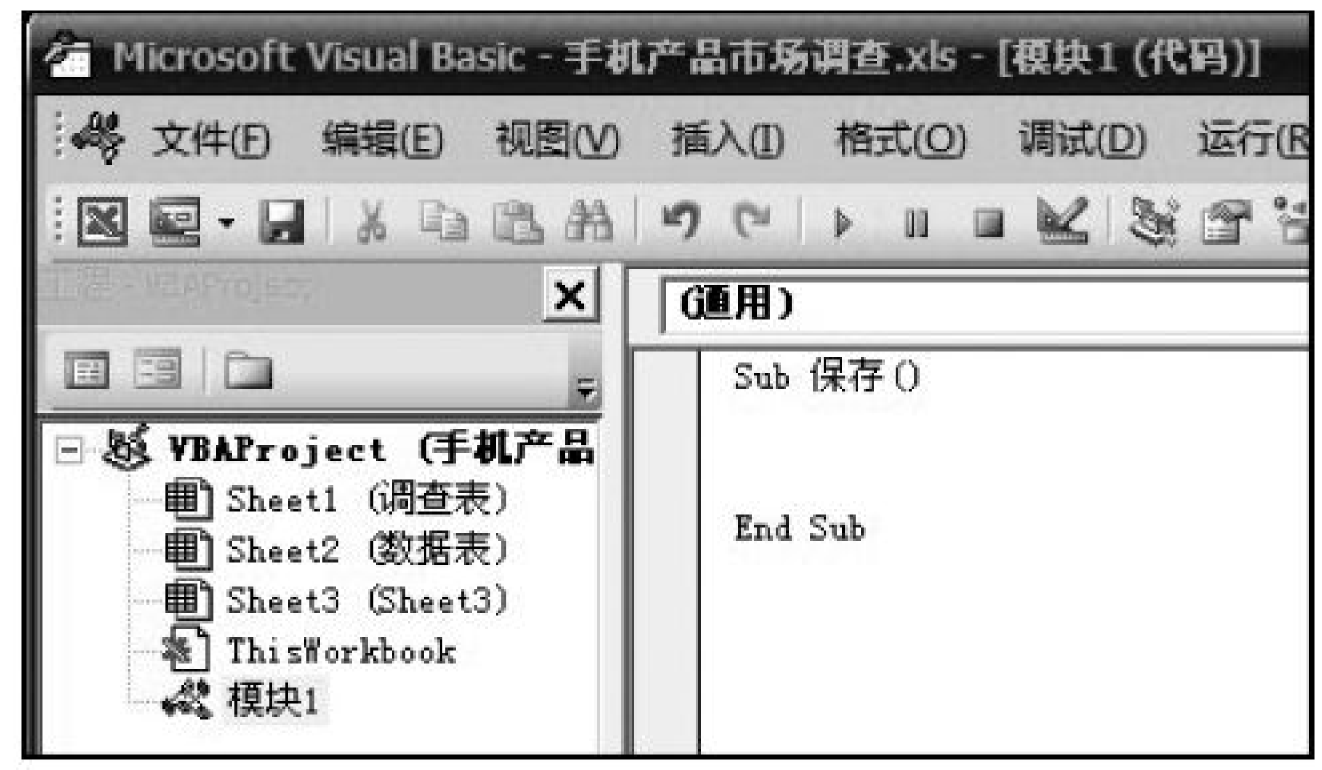


图 3-12 VB 编辑器窗口

这时再单击刚才打开的 VB 编辑器窗口，右侧代码区域会多出如下代码：

```
Sub 保存()
      Sheets("数据表").Select
      Range("A7").Select
      Selection.EntireRow.Insert
      Range("B3:J3").Select
      Selection.Copy
      Range("B7").Select
      ActiveSheet.Paste
   Sheets("调查表").Select
      ActiveWorkbook.Save
   End Sub
```

将这些代码“复制”，右键单击左侧的“模块 1”，选择“删除”选项；再右单击左侧的 VBAProject，选择插入→模块，在右侧的代码区将刚才的代码“粘贴”上。

6. 保存退出

右键单击“回答完成”按钮，选择“指定宏”选项，弹出的窗口中选择宏名为“保存”，再单击“确定”按钮。

此时，我们可以试着将前面的题目做完，然后单击“回答完成”按钮，此时打开“调查结果表”我们会发现，刚才我们的题目结果已经自动保存到调查结果表当中。

七、问卷的预评估

一般来讲，自主设计的问卷在设计完成后，并不能直接进行问卷调查收集数据，而是要先通过预调查，进行问卷的预评估，以保证问卷具有较高的可靠性和有效性，进而形成正式问卷。因此，在形成正式问卷之前，应当对问卷进行试测，并对试测结果进行信度和效度分析，根据分析结果筛选问卷题项，调整问卷结构，从而提升问卷的信度和效度。

问卷测验中测量误差通常来源于两个方面：一是产生于问卷测验过程中的误差，称为测量误差(measurement error)，也称为随机误差(random error)；二是由问卷的结构质

量造成的误差,称为系统误差(system error)。问卷的效度涉及问卷的准确性和有效性,它反映了对问卷的系统误差的控制程度。信度分析涉及问卷测验结果的一致性和稳定性,其目的是如何控制和减少随机误差。

(一) 问卷的效度评估

1. 效度的定义

效度(validity)即有效性,通常是指问卷的有效性和正确性,亦即问卷能够测量出其所欲测量目的的程度。其包括两个方面的含义:一是问卷测验的目的;二是问卷对测量目标的测量精确度和真实性。

在测量理论中,效度被定义为在一系列测量中,与测量目的有关的真变异数(有效变异)与总变异数之比:

$$r_{xy}^2 = \frac{S_v^2}{S_x^2}$$

式中,r_{xy}^2 为测量的效度系数;S_v^2 为有效变异数,S_x^2 为总变异数。

效度是问卷调查研究中最重要的特征,问卷调查的目的就是要获得高效度的测量与结论,效度越高表示该问卷测验的结果所能代表要测验的行为的真实度越高,越能够达到问卷测验目的,该问卷才正确而有效。对于一个标准测验来说,效度比信度更为重要。测验首先要保证能如实地测量出所要测量的东西,否则,这种测量将是没有意义的。

2. 效度的分类及测度

效度分为三种类型:内容效度(face validity)、准则效度(criterion validity)和结构效度(construct validity)。

1) 内容效度

内容效度又称表面效度或逻辑效度,它是指所设计的题项能否代表所要测量的内容或主题。内容效度常以题目分布的合理性来判断,属于命题的逻辑分析,所以,内容效度也称为"逻辑效度(logical validity)""内在效度(intrinsic validity)""循环效度(circular validity)"。例如,以考查学习成绩为目的的调查来说,调查的内容是指教学大纲所规定的全部教材,起预期反应的是指学生学习这些教材所产生的行为变化,如对教材的记忆、理解和应用。

在编制问卷时,内容效度是一个相当复杂的问题,如教师编制成绩调查问卷,其目的是了解学生在某一学科或专题上对知识的掌握情况,若条件允许,应该对大纲规定的所有内容进行全面考试,这显然是行不通的,只能从这一范围总体内容中选取有代表性题目(样本),组成测验,根据测验分数推论学生对该范围总体知识的掌握。若测验题目较好地代表了这个知识范围,则推论是有效的,即测验的内容效度高;若选题有偏差,则推论是无效的,即测验的内容效度低。另外,测验题目所引起的被试反应(是一个样本),若能代表其对本学科或专题的全部行为反应,也说明该测验是有效的;反之,则是无效的。因此,一个测验要有较高的内容效度应具备如下两个条件:一是要有定义好的内容范围;二是测验题目取样应有代表性(对所界定的内容范围而言)。

对内容效度常采用逻辑分析与统计分析相结合的方法进行评价。逻辑分析一般由研

究者或专家评判所选题项是否“看上去”符合测量的目的和要求。统计分析主要采用单项与总和相关分析法获得评价结果，即计算每个题项得分与题项总分的相关系数，根据相关是否显著判断是否有效。若问卷中有反意题项，应将其逆向处理后再计算总分。

2）准则效度

准则效度又称为效标效度或预测效度。准则效度分析是根据已经得到确定的某种理论，选择一种指标或测量工具作为准则（效标），分析问卷题项与准则的联系，若二者相关显著，或者问卷题项对准则的不同取值、特性表现出显著差异，则为有效的题项。评价准则效度的方法是相关分析或差异显著性检验。在调查问卷的效度分析中，选择一个合适的准则往往十分困难，使这种方法的应用受到一定限制。

3）结构效度

结构效度是指测量结果体现出来的某种结构与测值之间的对应程度，即某问卷测验的实际得分能解释某一特质的程度。如果我们根据理论的假设结构，通过问卷测验得到答卷者实际分数，经统计检验，结果表明问卷能有效解释答卷者该项特质，则说此问卷具有良好的结构效度。

结构效度分析所采用的方法是因子分析。效度分析最理想的方法是利用因子分析测量问卷或整个问卷的结构效度。因子分析的主要功能是从问卷全部变量（题项）中提取一些公因子，各公因子分别与某一群特定变量高度关联，这些公因子即代表了问卷的基本架构。透过因子分析可以考察问卷是否能够测量出研究者设计问卷时假设的某种结构。在因子分析的结果中，用于评价结构效度的主要指标有累积贡献率、共同度和因子负荷。累积贡献率反映公因子对问卷或问卷的累积有效程度，共同度反映由公因子解释原变量的有效程度，因子负荷反映原变量与某个公因子的相关程度。

3. 提高问卷效度的方法

（1）理论正确，解释清楚。问卷内容要适合问卷测验的目的，题目要清楚明了，易于理解，问卷的排列要由易到难，题目的难度和区分度要合适。

（2）操作规范以减少误差。

（3）控制系统误差。控制系统误差主要包括仪器不准，题目和指导语有暗示性，答案安排不当（被试可以猜测）等，控制这些因素可以降低系统误差，提高效度。

（4）样本适宜且要预防流失。重视问卷调查的回收率，样本容量一般不应低于30。

（5）适当增加问卷的长度。增加问卷的长度既可提高问卷的信度，也可以提高问卷的效度，但增加问卷的长度对信度的影响大于对效度的影响。

（6）排除无关因素干扰。认清并排除足以混淆或威胁结论的无关干扰变量。

（二）问卷的信度评估

1. 信度的定义

信度（reliability）即可靠性，它是指采用同样的方法对同一对象重复测量时所得结果的一致性程度。其包含两个层面的含义，一是相同的个体在不同时间，以相同的问卷测验，或以复本测验，或在不同的情景下测验，是否能得到相同的结果，即问卷测验结果是否随时间和地点等因素而变化；二是能否减少随机误差对问卷测验结果的影响，从而能够反

映问卷所要测量的真实情况,即问卷测验结果是否具有稳定性、可靠性和可预测性。一份好的问卷必须是稳定可靠的,且多次问卷测验结果应前后一致。

信度是用估计测量误差大小的尺度,来说明问卷测验结果中测量误差所占的比率。在测量理论中,信度被定义为信度可定义为真实分数(true score)的方差与测验实得分数(observed score)的方差之比:

$$r_{XX}=\frac{S_T^2}{S_X^2}\quad 或 \quad r_{XX}=1-\frac{S_E^2}{S_X^2}$$

其中,r_{XX} 为效度系数;S_T^2 为真实分数的方差;S_X^2 为实得分数的方差;S_E^2 为误差的方差。

当实得分数变异可以全部由真实分数的变异解释时,测验误差就是 0,这时问卷测验的信度为 1。

2. 信度的分类

信度指标多以相关系数来表示,大致可分为三类:稳定系数(跨时间的一致性)、等值系数(跨形式的一致性)和内在一致性系数(跨项目的一致性)。

若以信度系数来表示信度的大小,信度系数越大,表示测量的可信程度越大。一般认为,0.60~0.65(最好不要);0.65~0.70(最小可接受值);0.70~0.80(相当好);0.80~0.90(非常好)。由此,一份信度系数好的问卷或问卷最好在 0.80 以上,对于社科类调查问卷,0.70~0.80 还算是可以接受的范围;分问卷最好在 0.70 以上,0.60~0.70 可以接受。若分问卷的内部一致性系数在 0.60 以下或者总问卷的信度系数在 0.80 以下,应考虑重新修订问卷或增删题项。

3. 信度的测度

1)重测信度法

重测信度法是用同样的问卷对同一组被调查者间隔一定时间重复施测,计算两次施测结果的相关系数。显然,重测信度属于稳定系数。重测信度法特别适用于事实式问卷,如性别、出生年月等在两次施测中不应有任何差异,大多数被调查者的兴趣、爱好、习惯等在短时间内也不会有十分明显的变化。如果没有突发事件导致被调查者的态度、意见突变,这种方法也适用于态度、意见式问卷。由于重测信度法需要对同一样本试测两次,被调查者容易受到各种事件、活动和他人的影响,而且间隔时间长短也有一定限制,因此在实施中有一定困难。

2)复本信度法

复本信度法是让同一组被调查者一次填答两份问卷复本,计算两个复本的相关系数。复本信度属于等值系数。复本信度法要求两个复本除表述模式不同外,在内容、格式、难度和对应题项的提问方向等方面要完全一致,而在实际调查中,很难使调查问卷达到这种要求,因此采用这种方法者较少。

3)折半信度法

折半信度法是将调查项目分为两半,计算两半得分的相关系数,进而估计整个问卷的信度。折半信度属于内在一致性系数,测量的是两半题项得分间的一致性。这种方法一般不适用于事实式问卷(如年龄与性别无法相比),常用于态度、意见式问卷的信度分析。在问卷调查中,态度测量最常见的形式是 5 级李克特(Likert)问卷。进行折半信度分析

时，如果问卷中含有反意题项，应先将反意题项的得分做逆向处理，以确保各题项得分方向的一致性，然后将全部题项按奇偶或前后分为尽可能相等的两半，计算二者的相关系数（rhh，即半个问卷的信度系数），最后用斯皮尔曼-布朗（Spearman-Brown）公式：ru＝2rhh/(1＋rhh)求出整个问卷的信度系数(ru)。

（4）α 信度系数法。

Cronbach α 信度系数是目前最常用的信度系数，其公式为

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} S_i^2}{S_X^2}\right) \tag{3-3}$$

式中，n 为问卷中题项的总数；S_i^2 为答卷者在第 i 题得分的方差；S_X^2 为全部题项总得分的方差。从式(3-3)中可以看出，α 系数评价的是问卷中各题项得分间的一致性，属于内在一致性系数。这种方法适用于态度、意见式问卷的信度分析。一般问卷的 α 系数在 0.8 以上，社会科学类问卷 α 系数在 0.7 以上，该问卷才具有使用价值。

4. 提高问卷信度的方法

问卷的信度越高，受到人、时、地、物的干扰就越低，其所能反映事实或让人相信的程度就越高，因此在问卷实施前如何有效提高信度是问卷测验成败的关键。通常提高问卷信度的方法有以下几种。

（1）适当延长测验的长度。问卷的长度主要指问卷所包含的题目多少。对一个问卷来说，问卷的题目越少，得分越容易受偶然因素的影响，故问卷的信度越低。反之如果问卷题目较多，即问卷长度延长，扩大了被试得分范围，可在一定程度上排除偶然因素的影响，从而提高问卷信度。但是问卷信度的增加并不是等比例提高信度系数。当信度系数较小时，延长问卷长度，信度系数增加较大；当信度系数已经较大时，延长问卷长度对信度系数的影响就较小了。而且在延长问卷长度时，还需考虑其他因素的影响，如被试在回答问题时是否疲倦或产生厌烦情绪，是否节省时间、物力和财力，题项是否符合问卷目的等。

（2）问卷的难度要适中。难度即问卷的难易程度，当问卷难度太大时，被试得分普遍太低，呈负偏态分布；当问卷难度太小时，被试得分普遍较高，呈正偏态分布。太难太易的问卷都使被试得分差异减小，使实得分数方差减小，从而降低问卷信度。

（3）问卷的内容尽量同质。性质相同的问卷内容，对被试也要求相同的能力、知识和技能；而内容不同质的问卷，则要求被试不同的能力、知识和技能。因而为了提高问卷信度，问卷内容应尽量同质。

（4）问卷的时间要充分。对某一问卷而言，应保证绝大多数被试者在规定时间内完成问卷；否则，如果被试者不能从容回答所有问题，就不能反映被试者的真实水平。

（5）问卷的程序要统一。问卷的程序要统一包括问卷的题目统一，指导语、回答问题的方式、分收试卷的方法、问卷时间等都要统一。

（6）评分要客观。评分是否客观对问卷信度有直接的影响。对于客观性题目，评分标准明确，评分容易做到客观；但对于主观性题目，受评分者影响较大，不易做到客观。为了尽可能客观评分，应制定明确而易掌握的评分标准，尽量做到一卷多评，或一人只评一题等。

总之,为了提升问卷设计的质量,进而提升整个研究的价值,问卷的信度和效度分析非常重要,是研究过程中必不可少的重要环节。

第五节 调查数据的审核与插补

一、统计调查资料的审核

(一) 审核的含义

统计调查资料的审核是应用各种检查规则来辨别缺失、无效或不一致的录入。在对统计调查数据进行整理时,首先要进行审核。审核的目的是更好地了解调查的过程及调查数据,以确保调查数据的完整、准确与一致。审核工作贯穿于整个调查过程,从访问员在调查现场的简单初步的检查到数据处理阶段数据录入之后由计算机程序进行的复杂校验,为进一步整理与分析打下基础。

(二) 审核的内容和方法

从不同渠道调查取得的统计数据,在审核的内容和方法上都有所不同。

1. 直接调查取得的原始数据的审核

对于直接调查取得的原始数据应主要从完整性和准确性两个方面去审核。

(1) 完整性审核。检查应调查的单位或个体是否有遗漏,所有的调查项目或指标是否填写齐全。

(2) 准确性审核。检查数据是否真实反映客观实际情况,内容是否符合实际;检查数据是否有错误,计算是否正确等。

审核数据准确性的方法主要有逻辑检查和计算检查。

① 逻辑检查。逻辑检查是从定性角度,审核数据是否符合逻辑,内容是否合理,各项目或数字之间有无相互矛盾的现象。比如中学文化程度的人所填的职业是大学教师,对于这种违背逻辑的项目应查明原因,如果错了应及时予以纠正。此法主要用于对定类(列名)和定序(顺序)数据的审核。

② 计算检查。计算检查是检查调查表中的各项数据在计算结果和计算方法上有无错误。比如各分项数据之和是否等于相应的合计数,各结构比例之和是否等于 1 或 100%,出现在不同表格上的同一指标数值是否相同等。此法主要用于对定距和定比数据的审核。

2. 间接取得的二手数据的审核

对于间接取得的二手数据,除了从完整性和准确性两个方面进行审核外,还应着重审核数据的适用性和时效性。二手数据可以来自多种渠道,有些数据可能是为特定目的通过专门调查而取得的,或者已经按特定目的的需要做了加工整理,对于使用者来说,应该做到以下几点。

(1) 适用性审核。弄清楚数据的来源、数据的口径以及有关的背景材料;确定数据是否符合自己分析研究的需要。

(2) 时效性审核。尽可能使用最新的数据，如果数据过于滞后，可能失去了研究的意义。一般来说，应使用最新的数据。

(3) 确认是否有必要做进一步的加工整理。

从数据的表现性质来看，进行数据审核的内容主要有有效性审核、一致性审核和分布审核。

(1) 有效性审核。有效性审核主要是检查被调查者回答语句的语法是否正确，检查包括是否在规定填数字的地方填上了非数字的字符以及问卷中的回答是否有缺失等各种错误。

(2) 一致性审核。一致性审核主要是检查不同问题之间的关系是否正确。一致性审核可以基于不同问题或同一问题的不同部分之间的结构关系、逻辑性和合法性来进行。

(3) 分布审核。分布审核主要是试图通过数据的分布，来辨识记录是否远远脱离分布的正常范围，即是否为离群值。又称为离群值的检测，第三个标题将详细讲解。分布审核主要是用来发现和确认可疑的记录。

对审核过程中发现的错误应尽可能予以纠正。

二、调查数据的插补

(一) 插补的含义

插补就是解决在审核过程中辨别出来的数据缺失、无效与不一致等问题的过程。

(二) 插补方法

插补方法分为两类，即确定性插补和随机插补。确定性插补，对于特定的被调查者的数据，可能插补值只有一个。随机插补含有随机因素。换句话说，如果对同一组数据进行多次插补，对于确定性的插补每次都是相同的值；而对于随机插补，每次得出的值可能会不一样。

确定性插补的具体方法有：推理插补、均值插补、比率或回归插补、热平台插补、冷平台插补、最近邻插补。每一种确定性的插补方法都对应着一种随机插补方法。用确定性的方法得出一个插补值，加上从某个适宜的分布或模型产生出的一个残差作为最后插补值，就称为随机插补。

1. 推理插补

将缺失的或不一致的数据通过推断来确定。这种推理是根据问卷上其他回答项的模式来进行的。如三项之和为100，空着一项。

2. 均值插补

用插补类的均值代替缺失或不一致的值。如假定在一份住房调查的问卷中，公寓月租金的值缺失，则利用同插补类中正确填报租金的问卷计算其平均值(插补类由居住在同一地区的被调查者所填的需要进行插补处理的问卷构成)。

3. 比率或回归插补

比率或回归插补是使用辅助信息及其他记录中的有效回答建立一个比率或回归模

型,该模型表明了两个或多个变量之间的关系。

$$若\ y_i = Rx_i + \varepsilon_i \quad 则\ y_{插i} = \frac{\bar{y}}{\bar{x}} x_i$$

式中,$\bar{x}$ 为插补类中记录 x 的均值;$\bar{y}$ 为插补类中记录 y 的均值。

插补值的精度在很大程度上取决于要插补的变量与已知的变量是否存在密切关系,取决于所运用的数学计算以及这种计算是否严格限制在一个插补类中或是全部数据集中。

4. 热平台插补

热平台插补是使用同一插补类中的供者记录的信息来代替一个相似的受者记录中缺失的或不一致数据的插补方法。热平台插补分为序贯热平台插补和随机热平台插补。

在序贯热平台插补方法中,数据在插补类中是按某种顺序排列进行处理的,插补就是用这个序列需要插补数据前面某一个有效的回答单元的数据来代替缺失的数值。如果每次都使用相同的排序及选取方法,就是确定性的序贯热平台法。

而随机热平台插补,供者是在插补类中随机选出的。例如,我们希望插补被调查者的吸烟状况。有三种可能的回答:吸烟、不吸烟及以前吸过但现在已戒烟。要找到一个相似的供者记录,我们基于年龄段和性别产生插补类(因为我们认为吸烟状况和一个人的年龄和性别有关)。假设要做插补处理的记录是女性,在 15～24 岁这个年龄组中。要找到一组供者,我们看所有在同一年龄组的女性被调查者(她们回答了各自的吸烟状况)。要从这一组中选择一供者,可以随机地选择一个(随机热平台法)或者按某种顺序的供者清单,从中选一个(序贯热平台法)。

热平台插补的优点:因为供者与受者具有相似的特征,因此插补出的数值应相当准确,可以保持数据的原始分布形式。

热平台插补的缺点:序贯热平台插补经常导致同一个供者的多次使用,从而降低(理解为增大)抽样误差;有时很难找到一个合适的供者,因为建立插补类需要足够的辅助信息或者至少需要部分回答。

为了保证总能找到一个供者记录,可以用多层次热平台插补。当在最初的插补类中找不到供者记录时,这些类就按照层次结构合并,直到找到一个供者。

5. 冷平台插补

与热平台插补不同之处在于热平台插补使用当前调查的供者,而冷平台插补则使用前期的或普查中的供者资料。

6. 最近邻插补

最近邻插补像热平台插补一样,也是基于匹配变量选择一个供者记录。但是,用这种方法,目的不是非要找出一个和受者记录在匹配变量上完全相同的供者记录,而是插补类中按匹配变量找到和受者记录最接近的供者记录——找到距离最近的值。“最近”是通过两个观测对象之间的距离来定义的,两个观测对象之间的距离是由辅助数据计算的。当匹配变量的纲不同时(如货币和土地面积),运用最近邻插补时应格外小心,此时应该先将匹配变量尺度标准化;然后再进行插补。

（三）几个技术问题

1. 确定哪些值需要插补

我们并不主张对所有审核失效的数据都进行插补，应该通过变更尽可能少的数据项（字段），以使每条记录都满足审核规则的要求。

例如，假设某份问卷中关于一位被调查者的背景资料是：受教育程度（大学），婚姻状况（已婚），性别（女），年龄（10 岁）。显然这条记录中，年龄与婚姻状况、年龄与受教育程度是不符合审核规则的。为了纠正审核失效，可以同时调整婚姻状况和受教育程度，也可以只对年龄做调整。我们倾向于后者。

2. 怎样为受者找到供者记录

选择匹配变量时必须注意要使匹配变量与插补类中需要插补的变量密切相关，然后再用这些匹配变量去查找供者记录。

3. 是不是某个受者记录中的所有字段都应该用一个供者来插补

在有些情况下，因为用同一个记录中的所有字段进行插补能保持变量的联合分布，所以用同一个供者来插补某个受者记录中的所有字段。例如，在一项劳动力调查中，如果职业和个人收入都需要进行插补，那么根据相同的供者记录来插补受者记录中的这两个缺失或无效数据就具有明显的优点，因为这两个变量之间存在相互关系。这种单一供者热平台插补的方法的另一个优点是，插补时考虑到了审核失效的问题，从而大大减少了以后步骤中的审核失效。

但是，如果有很多变量需要插补，对插补一个字段合适的匹配变量（或者使用来建立热平台或冷平台插补中插补类的变量），对另一个字段可能不合适，特别是当与需要插补的变量无关的时候。考虑一项多目的的健康调查，在这项调查中，被调查者的身高和每天的吸烟量是需要插补的两个变量。在这种情况下，每个需要插补的变量用一组不同的匹配变量就比较合适。但如果确定太多的匹配变量，就会有找不到合适供者的危险。

在通常情况下，对于热平台插补，整个插补过程分成几个阶段，每个阶段插补几个变量。其结果是，在完成一个不完善的记录的插补时，可能要涉及几个供者。如果情况真是这样，需要插补的某个“关键”变量可以继续为后面阶段组建插补类所用，从而保证插补记录内在的完整性。

4. 一个供者记录能用来插补一个以上的受者记录吗

如果几个受者记录都由同一个供者记录来插补，对最终调查估计产生较大的偏倚。

三、离群值的检测

（一）离群值的含义

离群值被定义为一个观测值或一组观测值，它们看起来与数据集中的其他观测值不一致。

(二)离群值的检测方法

1. $d_i=\dfrac{|y_i-m|}{s}$

式中,$y_i=y_1,y_2,\cdots,y_n$ 为样本数据;m 为数据集中趋势的度量;s 为离散趋势的度量。

若 d_i 大于预先确定的偏离值,那么该观测值就被认为是一个离群值。

2. $(m-c_ls,m+c_us)$

式中,c_l、c_u 分别为预先确定的下限和上限值。如果总体是偏态的,c_l 和 c_u 就不相等。落在这个区间之外的观测值被认为是离群值。

3. $(q_{0.5}-c_lh_l,q_{0.5}+c_uh_u)$

式中,$q_{0.25}$ 为第一四分位数;$q_{0.5}$ 为第二四分位数(即中位数);$q_{0.75}$ 为第三四分位数,$h_l=q_{0.5}-q_{0.25}$;$h_u=q_{0.75}-q_{0.5}$;c_l 和 c_u 可以通过以前的数据或基于过去的经验来确定。

任何落在这个区间之外的观测值都被认为是一个离群值。

如果数据成群偏在一边,样本均值就会偏向离群值,样本方差也会由于离群值而显著增长,此时可用四分位数法。用中位数度量数据的集中趋势、四分位数间距度量数据的离散程度。

四、离群值的处理

在手工审核系统中,对离群值进行检查,如果确认是错误,就要回访并校正。

在自动审核系统中,离群值经常要进行插补处理。

在审核时没有进行处理的离群值可以在估计的时候处理。

简单地忽略未经处理的离群值会影响估计的效果,并导致估计量的方差增大;给离群值赋予 1 或 0 的权数,估计结果发生偏倚。离群值处理的目的就是在不引入较大偏倚的前提下,尽量减少离群值对估计量抽样误差的影响。

估计时有三种方法可以处理离群值。

(一)改变数值

处理极值的一种方法是缩尾化,这种方法首先要将样本数据从小到大依次排序;然后再按下面的步骤计算。

在简单随机抽样中,总体总量 Y 的无偏估计公式如下:

$$\hat{Y}=\frac{N}{n}\sum y_i$$

对于缩尾化,假设 $y_i(i=1,2,3,\cdots,n)$是将样本观测值按从小到大的顺序重新排列得到的有序样本。若样本中前 k 个最大值(有序样本值中的 $y_{n-k+1},\cdots,y_n$)被认为是离群值,则单侧 k 次缩尾估计量就可通过用第$(n-k)$个最大值 y_{n-k}代替所有离群值来定义,即

$$\hat{Y}_w=\frac{N}{n}\left(\sum_{i=1}^{n-k}y_i+ky_{n-k}\right)$$

（二）调整权数

处理离群值的另一种方法是，降低离群值的权数使它们的影响变小。

若赋予离群值的抽样权数为 1 或 0，这样做对估计的影响太大，特别是对偏态总体的估计，会使估计出现严重的偏倚——通常是低估。

例如，如果两个大公司的零售额占总行业的绝大比重，而其中一个公司的零售额被确定为离群值，若从估计中剔除这个离群值，显然就会严重低估整个行业的零售额。

（三）使用稳健估计量

经典的估计理论中，总体参数的估计是基于某种分布的假设。通常，假定估计量服从正态分布，样本均值和方差估计量在正态性假定下最优。但是，这些估计量对离群值非常敏感。稳健估计量则能克服这种局限性，因为它对分布的假定不太敏感。比如，中位数比均值更稳健，四分位数间距比方差估计量更稳健。

第六节　统计调查误差

一、统计调查误差的概念

统计调查误差，是指调查结果与所调查现象的真实结果之间的离差。在多数情况下，调查误差是一个相对的概念。如果我们能准确地计算出调查误差，就意味着我们知道现象的真值，而我们知道现象的真值，研究调查误差也就失去了意义。所以，我们研究调查误差的目的不是如何去定义和测量它，而是分析其产生的原因，进而在统计调查工作中尽量避免或减少误差。

二、统计调查误差的种类

一般可以按照误差产生的原因将统计调查误差分为登记性误差和代表性误差两类。

（一）登记性误差

登记性误差是指在调查过程中，由于各种主观因素的影响而引起的技术性、登记性误差以及责任性误差等。具体而言，调查者在调查或整理过程中重复登记、遗漏登记、错误登记、错误记录或丢失数据等，都会造成调查误差；被调查者在接受调查时错误理解、记忆不清、有意隐瞒或夸大事实等，也会造成调查误差。登记性误差在全面调查和非全面调查中都可能会产生。

（二）代表性误差

代表性误差是指由于样本的代表性不足而造成样本指标与总体指标之间可能存在的误差。代表性误差是用部分总体单位来推断总体时产生的误差。由于总体单位间存在差异，任何一部分总体单位对总体而言总是没有完全的代表性，用它们来推断总体的特征必

然会因代表性问题而产生误差。

全面调查中不存在代表性误差,只有用于推断总体的非全面调查才会有代表性误差,比如抽样调查等。

三、统计调查误差的避免与控制

(一)登记性误差的避免

登记性误差是人为造成的,因此有可能加以避免和消除。针对具体产生的原因可以采取以下措施。

(1)加强对调查人员的业务培训,增强其责任心,提高其业务素质。

(2)合理设计调查表,不使被调查者产生误解。

(3)采用合适的调查组织方式,尽量避免人为因素的干扰。

(二)代表性误差的控制

只要调查单位不是全部总体单位,代表性误差就不可避免,但我们可以加以控制。通常的做法是按随机原则抽取样本、适当扩大样本容量、采用合适的抽样组织方式等。

思　考　题

1. 什么是统计调查?统计调查有哪些基本要求?
2. 统计调查的种类有哪些?
3. 简述普查和抽样调查的特点。
4. 简述典型调查和重点调查的区别与联系。
5. 简述调查方案的内容。
6. 什么是问卷?它由哪几部分组成?
7. 设计问卷的提问项目时应注意的问题有哪些?
8. 问卷中问题顺序的设计应注意哪些问题?
9. 什么是统计调查误差?产生误差的原因有哪些?
10. 简述审核与插补的含义与方法。

练　习　题

1. 假定我们要调查某城市居民的住房状况,请写出你的调查思路。

2. 某家用电器生产厂家想通过市场调查了解以下问题:

企业产品的知名度,产品的市场占有率,用户对产品质量的评价及满意度。

要求:(1)请你设计出一份调查方案。

(2)你认为采用哪种调查方法比较合适?

(3)利用 Excel 设计一份调查问卷,能自动保存调查结果数据。

第四章

统计整理

【统计知识 ABC】

[A]贡献率

贡献率是分析经济效益的一个指标。它是指有效或有用成果数量与资源消耗及占用量之比,即产出量与投入量之比,或所得量与所费量之比。计算公式为

贡献率(%) = 贡献量(产出量,所得量)/ 投入量(消耗量,占用量) ×100%

贡献率也用于分析经济增长中各因素作用大小的程度。

计算方法是:贡献率(%)=某因素增加量(增量或增长程度)/总增加量(总增量或增长程度)×100%

上式实际上是指某因素的增长量(程度)占总增长量(程度)的比重。

例如,国内(地区)生产总值中第一、第二、第三产业贡献率为:第一、第二、第三产业增加值增量分别与国内(地区)生产总值增量之比。

(资料来源:http://baike.baidu.com/view/43725.html.)

[B]拉动力

拉动力是指在总的经济增长率中带动的百分点数。如产业部门拉动力是指在 GDP 增长中各产业部门拉动的百分点数,需求拉动力是指在 GDP 增长中各需求构成(消费需求、投资需求、净出口需求)拉动的百分点数。其计算公式为

拉动力(%)=贡献率(%)×GDP 增长率(%)

(资料来源:http://baike.baidu.com/view/857522.html.)

[C]人口老龄化

人口老龄化是指总人口中因年轻人口数量减少、年长人口数量增加而导致的老年人口比例相应增长的动态。两个含义:一是指老年人口相对增多,在总人口中所占比例不断上升的过程;二是指社会人口结构呈现老年状态,进入老龄化社会。国际上通常看法是,当一个国家或地区 60 岁以上老年人口占人口总数的 10%,或 65 岁以上老年人口占人口总数的 7%,即意味着这个国家或地区的人口处于老龄化社会。

第六次人口普查数据显示:60 岁及以上人口占 13.26%,比 2000 年人口普查上升 2.93 个百分点,其中 65 岁及以上人口占 8.87%,比 2000 年人口普查上升 1.91 个百分点。

由此判断,中国大约在 2000 年开始就进入了老龄化社会。

(资料来源:http://baike.baidu.com/view/109749.html.)

统计整理是统计工作过程的中间环节,它既是统计调查的继续,又是后续进行统计分析的基础,在整个统计过程中起承上启下的作用。

第一节　统计整理概述

一、统计整理的概念和意义

统计整理就是根据统计研究的目的,将统计调查所采集到的原始资料进行科学的分类和汇总,或对已初步加工的资料进行再加工,使之系统化、条理化,并以某种特定的方式展示出来,成为能够反映现象总体特征的综合资料的工作过程。

统计调查采集到的原始资料都是个别的、分散的,根据这样的资料,人们难以认识和分析社会经济现象的本质特征及数量变化的规律性。例如,在全国人口普查中所收集到的每个居民的资料,只能说明每个居民的性别、年龄、民族、文化程度、职业等具体情况。为了说明全国人口总体的综合特征,如全国的人口总数、性别比例、年龄构成、民族构成、职业构成等,必须对原始资料进行分组、汇总,形成系统化的资料,并以图表的形式展示出来,才能反映全国人口的综合特征,进而通过比较分析和研究,来确立我国人口发展的策略及经济发展的重点。

因此,统计整理是人们对社会经济现象从感性认识上升到理性认识的过渡阶段。统计整理结果的质量决定了统计资料的科学价值,也直接影响到统计分析的准确性和可靠性。

二、统计整理的内容

统计整理工作一般包括以下五个方面的内容。

(一) 统计调查资料的预处理

对调查资料进行分类或分组之前必须进行预处理,主要包括数据的审核、筛选、排序、分类汇总等。

(二) 统计分组

科学的分组是搞好统计整理的前提,统计分组就是依据一定的分组标志,将原始资料划分为若干组别的过程。通过分组,可以将总体中千差万别的总体单位划分为若干大类,达到“求同存异”的效果。

(三) 统计汇总

统计汇总是选择适当的汇总组织形式和技术方法,按分组要求对原始资料进行汇总,计算出各组及总体的单位总量和标志总量。统计汇总是统计整理的中心内容。

(四) 整理结果的显示

统计表和统计图是表达统计资料整理结果的重要形式。统计表和统计图将汇总得到

的各项指标数据按照表格所列的项目填列在相应的栏目内,可以更加简明、直观地反映被研究现象的发展趋势和内在规律。使用者可以根据研究目的的不同,编制形式各异的统计表和绘制形状不同的统计图。

(五) 统计资料的保存

编制好的统计台账和加工整理后的统计资料,必须妥善保管,不得损坏和遗失。

以上五个方面是相互衔接的,其中,统计分组是统计整理的基础,统计汇总是统计整理的中心内容,统计表和统计图是统计整理结果的表现形式。

第二节 统计调查资料的预处理

统计调查资料的预处理(statistical data pretreatment)是数据分组整理的先前步骤,内容包括调查数据的审核与插补(第三章已经介绍)、筛选、排序、分类汇总等过程。

一、统计调查资料的筛选

在对统计调查数据进行整理时,首先需要进行审核,以保证数据的质量。对审核中发现的错误应尽可能予以纠正。如果对发现的错误无法纠正,或者有些数据不符合调查的要求而又无法弥补时,就要对数据进行筛选。

数据筛选(filter)有两方面内容:一是将某些不符合要求的数据或有明显错误的数据予以剔除;二是将符合某种特定条件的数据筛选出来,不符合特定条件的数据予以剔除。数据筛选可借助于 Excel 自动完成,下面我们通过举例说明其具体做法。

【例 4-1】 图 4-1 所示为某高校人员的基本情况资料。试进行筛选操作。

	A	B	C	E	G	H	I	J	Q
1									
2	人员编号	姓名	性别	学历	职务系列	职称	婚姻状况	基本工资	实发工资
3	001	韩利生	男	硕士	教师系列	副教授	☑ 已婚	2 300	3 219.9
4	002	李明双	女	博士	教师系列	副教授	☐ 已婚	1 800	2 756.4
5	003	马善溪	男	大专	工程系列	高级工程师	☑ 已婚	2 400	3 649.2
6	004	马奎	男	大专	工程系列	高级工程师	☑ 已婚	2 400	3 649.2
7	005	邢利民	男	本科	教学管理系列	副研究员	☑ 已婚	2 200	3 141.6
8	006	李铁征	男	本科	教学管理系列	研究员	☐ 已婚	2 500	3 718.5
9	007	于海英	女	硕士	教师系列	副教授	☑ 已婚	1 900	2 933.7
10	008	刘长运	男	本科	教学管理系列	研究员	☑ 已婚	2 580	3 790.14
11	009	庄敏华	女	硕士	教师系列	副教授	☑ 已婚	1 850	2 867.55
12	010	李景义	女	大专	工程系列	工程师	☑ 已婚	2 040	3 097.32
13	011	刘荣旺	男	高中	工程系列	工程师	☑ 已婚	1 700	2 831.1
14	012	于西贝	女	大专	教学管理系列	副研究员	☑ 已婚	1 600	2 752.8
15	013	刘互助	男	大专	工程系列	高级工程师	☑ 已婚	2 500	3 727.5
16	014	李彤	女	博士	教师系列	教授	☑ 已婚	2 350	3 502.05
17	015	魏亚英	男	大专	工程系列	工程师	☑ 已婚	2 200	3 492.6
18	016	尹桂英	女	大专	工程系列	工程师	☑ 已婚	1 900	2 987.7
19	017	苏俊枫	男	本科	教学管理系列	副研究员	☑ 已婚	2 100	3 378.3
20	018	朱红光	男	本科	教学管理系列	副研究员	☑ 已婚	1 320	2 470.56

图 4-1 某高校人员的基本情况资料

1. 单字段条件筛选

如果筛选的条件只涉及一个标题字段,称为单字段条件筛选,比如筛选“实发工资”大于 3 000 元的员工。Excel 2013 中的操作步骤如下。

(1) 用鼠标选定标题行。

(2) 选择“数据”菜单选项,并单击“筛选”命令。

这时会在选定的标题行字段下出现下拉箭头,单击实发工资右侧的三角符,在弹出的菜单项中依次选择数字筛选→“大于”,弹出“自定义自动筛选方式”对话框,如图 4-2 所示。



图 4-2 “自定义自动筛选方式”对话框

要筛选出实发工资大于 3 000 元的员工,在大于右侧的框内输入“3 000”,单击“确定”按钮即可。如果筛选其他条件,直接在条件下拉列表框内选择即可。

2. 多字段条件筛选

如果筛选的条件涉及的标题字段多于一个,称为多字段条件筛选,比如筛选“实发工资”大于 3 000 元并且“基本工资”大于 2 000 元的员工,这是就需要使用 Excel 的“高级筛选”功能。Excel 2013 中的操作步骤如下。

(1) 建立条件区域。首先需要在数据清单上面至少留出三行作为条件区域,然后把数据清单中含有筛选值的数据列复制粘贴到条件区域的一个空行,最后在条件标志下面的一行中键入要匹配的条件。如图 4-3 所示。

(2) 用鼠标选定员工基本情况数据区域。

(3) 单击“数据”→筛选和排序按钮组中的“高级”,弹出“高级筛选”对话框,如图 4-4 所示。

	A	B	C	E	G	H	I	J	Q
1									
2					基本工资	实发工资			
3					>2 000	>3 000			
5	人员编号	姓名	性别	学历	职务系列	职称	婚姻状况	基本工资	实发工资
6	001	韩利生	男	硕士	教师系列	副教授	☑ 已婚	2 300	3 219.9
7	002	李明双	女	博士	教师系列	副教授	☐ 已婚	1 800	2 756.4
8	003	马善溪	男	大专	工程系列	高级工程师	☑ 已婚	2 400	3 649.2
9	004	马奎	男	大专	工程系列	高级工程师	☑ 已婚	2 400	3 649.2
10	005	邢利民	男	本科	教学管理系列	副研究员	☑ 已婚	2 200	3 141.6
11	006	李铁征	男	本科	教学管理系列	研究员	☐ 已婚	2 500	3 718.5
12	007	于海英	女	硕士	教师系列	副教授	☑ 已婚	1 900	2 933.7
13	008	刘长运	男	本科	教学管理系列	研究员	☑ 已婚	2 580	3 790.14
14	009	庄敏华	女	硕士	教师系列	副教授	☑ 已婚	1 850	2 867.55
15	010	李景义	女	大专	工程系列	工程师	☑ 已婚	2 040	3 097.32
16	011	刘荣旺	男	高中	工程系列	工程师	☑ 已婚	1 700	2 831.1
17	012	于西贝	女	大专	教学管理系列	副研究员	☑ 已婚	1 600	2 752.8

图 4-3 高级筛选



图 4-4 “高级筛选”对话框

(4) 列表区域自动显示我们选定的数据区域,只需在条件列表区输入筛选的条件区域,在本例中输入“$G$2:$H$3”,再单击“确定”按钮就可得到最终的筛选结果。

二、统计调查资料的排序

数据排序就是按一定顺序将数据排列，其目的是便于研究者通过浏览数据发现一些明显的特征或趋势，找到解决问题的线索。排序还有助于对数据检查、纠错，为重新分组或归类提供依据。在某些场合，排序本身就是分析的目的之一。

（一）定性数据的排序

定性数据的排序有以下两种。

（1）字母型数据。字母型数据的排序有升序降序之分，但习惯上用升序。

（2）汉字型数据。汉字型数据的排序可按汉字的首位拼音字母排列，也可按笔画排序，其中也有按笔画多少的升序降序之分。

比如中国北京第29届奥林匹克运动会各参赛国家的入场顺序，希腊是奥运会的发源地第一个入场，中国是东道主最后一个入场，其余两百多个国家的入场顺序就是按照该国家汉语名称的笔画顺序排序入场的。

（二）定量数据的排序

定量数据的排序有以下两种。

（1）递增排序。设一组数据为 $x_1, x_2, \cdots, x_n$，递增排序后可表示为：$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$。

（2）递减排序。（1）中的例子可表示为：$x_{(1)} > x_{(2)} > \cdots > x_{(n)}$。

无论是定性数据还是定量数据，其排序均可借助 Excel 完成。下面通过实例说明 Excel 2013 中进行数据排序的操作。

【例 4-2】 对图 4-1 中的数据按照实发工资的递增顺序进行排列。

操作步骤如下。

（1）单击标题行的任何一个单元格。

（2）选择“数据”菜单中的“排序”选项，出现“排序”对话框，如图 4-5 所示。

（3）在对话框中，单击“主要关键字”右边的下拉列表按钮，在字段下拉列表中选择主关键字段，如“实发工资”。

（4）选择“排序依据”选项，单击排序依据下方的下拉列表按钮，选择“数值”选项。

（5）选择“升序”还是“降序”选项，单击次序下方的下拉列表按钮，选择“升序”选项。

（6）还可以单击排序窗口上方的“添加条件”按钮，用上述同样的方法选择“次要关键字”“第三关键字”以及“排序依据”和“次序”选项，如图 4-6 所示。

（7）排除字段名行。因为字段名行不参加排序，所以数据清单中如果含有字段名行，应单击“数据包含标题”复选按钮将其排除。

（8）单击“确定”按钮，执行排序。

三、统计调查资料的分类汇总

在对数据进行预处理时，有时需要对某些字段按条件进行汇总，称为数据的分类汇

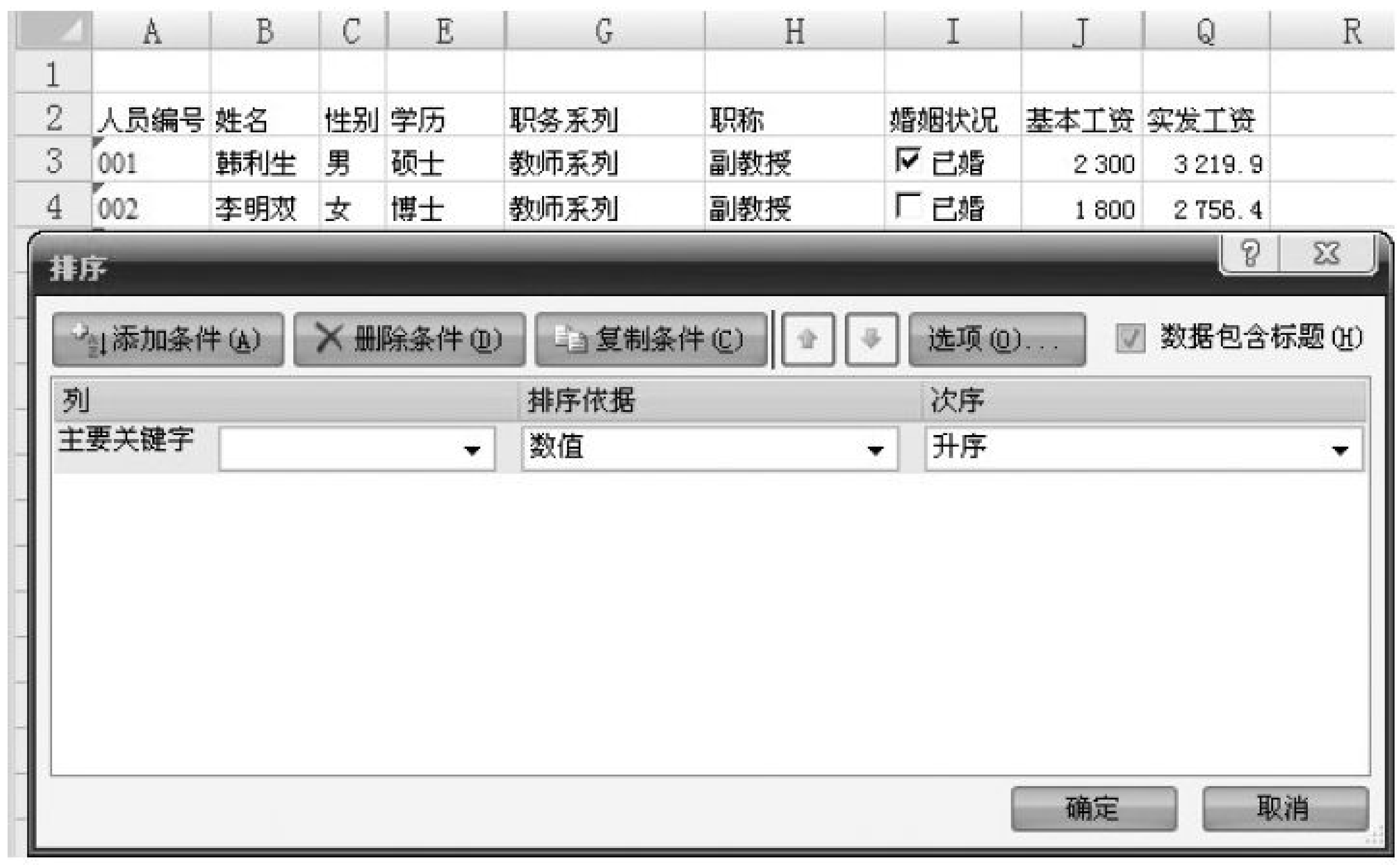


图 4-5 “排序”对话框



图 4-6 排序

总。如果只是针对一个字段进行分类汇总,称为单字段分类汇总;如果同时对两个及两个以上字段进行分类汇总称为多字段分类汇总。

分类汇总同样可以借助计算机完成。下面通过实例说明 Excel 2013 中分类汇总的操作。

(一) 单字段分类汇总

【例 4-3】 对图 4-1 中的数据按照职称进行分类汇总。

具体操作步骤如下。

(1) 用鼠标左键选择需要分类汇总的数据区域 A2:Q18。

(2) 按照“职称”进行排序。

(3) 选择“数据”菜单,单击“分类汇总”按钮,打开“分类汇总”对话框,如图 4-7 所示。

(4) 在“分类字段”的下拉式列表中,选择要进行分类的标题。对于本例,我们可以选择职称。在“汇总方式”的下拉式列表中,选择要进行汇总的方式。对于本例,我们可以选择求和。在“选择汇总项”中,选择要进行汇总的内容。对于本例,我们可以选择基本工



图 4-7　“分类汇总”对话框

资、实发工资。单击“确定”按钮，即可得到分类汇总的结果，如图 4-8 所示。



	A	B	C	E	G	H	I	J	Q
2	人员编号	姓名	性别	学历	职务系列	职称	婚姻状况	基本工资	实发工资
3	001	韩利生	男	硕士	教师系列	副教授	已婚	2 300	3 219.9
4	002	李明双	女	博士	教师系列	副教授	已婚	1 800	2 756.4
5	007	于海英	女	硕士	教师系列	副教授	已婚	1 900	2 933.7
6	009	庄敏华	女	硕士	教师系列	副教授	已婚	1 850	2 867.55
7						副教授 汇总			11 777.55
8	005	邢利民	男	本科	教学管理系列	副研究员	已婚	2 200	3 141.6
9	012	于西贝	女	大专	教学管理系列	副研究员	已婚	1 600	2 752.8
10	017	苏俊枫	男	本科	教学管理系列	副研究员	已婚	2 100	3 378.3
11	018	朱红光	男	本科	教学管理系列	副研究员	已婚	1 320	2 470.56
12						副研究员 汇总			11 743.26
13	003	马善溪	男	大专	工程系列	高级工程师	已婚	2 400	3 649.2
14	004	马董	男	大专	工程系列	高级工程师	已婚	2 400	3 649.2
15	013	刘互助	男	大专	工程系列	高级工程师	已婚	2 500	3 727.5
16						高级工程师 汇总			11 025.9
17	010	李景义	女	大专	工程系列	工程师	已婚	2 040	3 097.32
18	011	刘荣旺	男	高中	工程系列	工程师	已婚	1 700	2 831.1
19	015	魏亚英	男	大专	工程系列	工程师	已婚	2 200	3 492.6
20	016	尹桂英	女	大专	工程系列	工程师	已婚	1 900	2 987.7
21						工程师 汇总			12 408.72
22	014	李彤	女	博士	教师系列	教授	已婚	2 350	3 502.05
23						教授 汇总			3 502.05
24						总计			50 457.48

图 4-8　分类汇总的结果

（二）多字段分类汇总

【例 4-4】　对图 4-1 中的数据按照职称和性别分别对基本工资进行分类汇总。

按性别统计各职称的基本工资总额，建立一个交叉式的复合分组统计表，可使用数据透视表功能。其步骤如下。

（1）选择“插入”选项卡中的“数据透视表”命令，如图 4-9 所示，弹出“数据透视表”对话框，如图 4-10 所示。根据需要选择“数据源类型和数据源”以及数据透视表的位置。



图 4-9 “插入”选项卡中的“数据透视表”命令

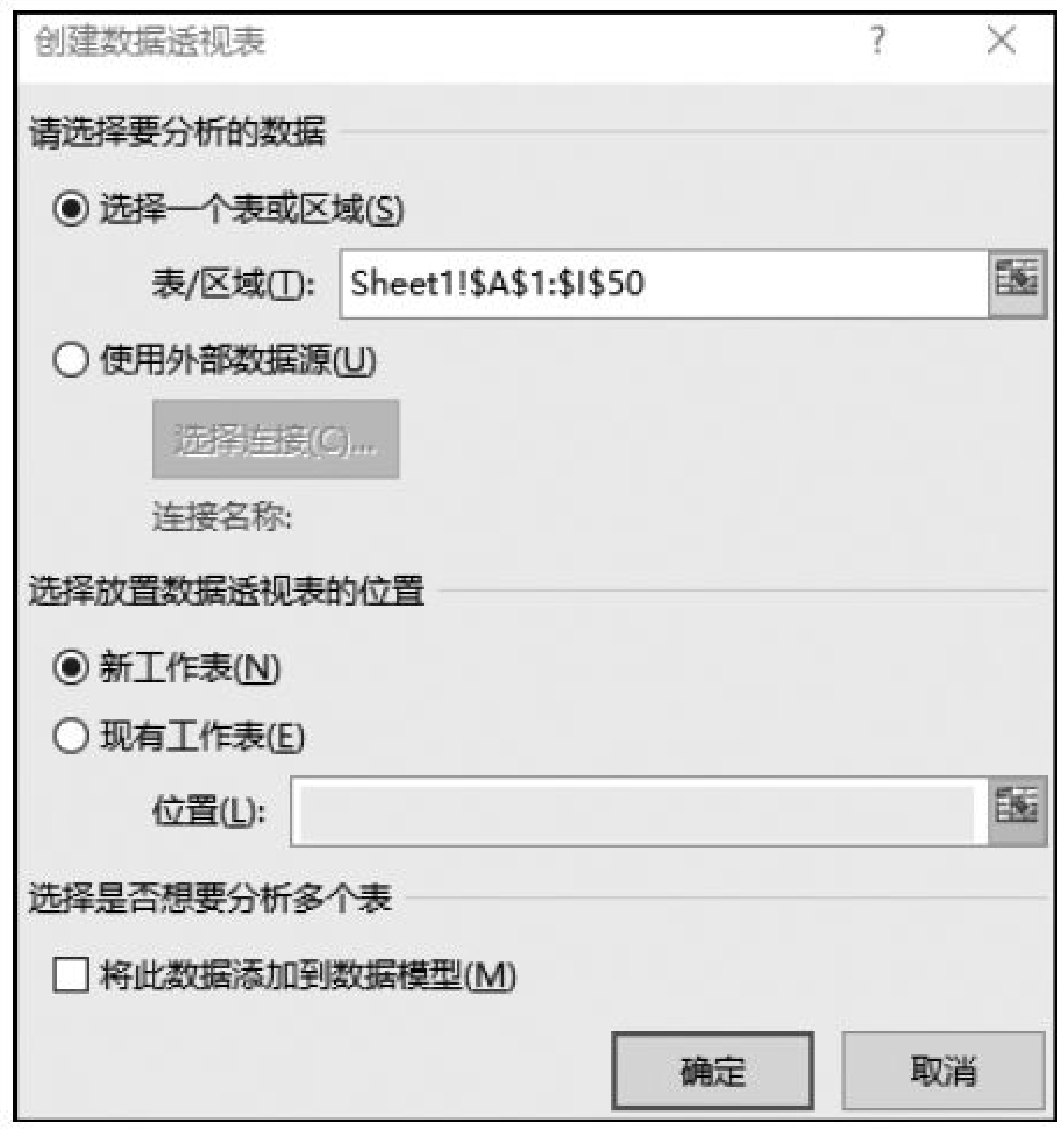


图 4-10 “数据透视表”对话框

(2) 单击“确定”按钮后,创建一新工作表,右侧显示“数据透视表字段”窗格,选定三个字段,Excel 自动添加到工作表中,如图 4-11 所示。

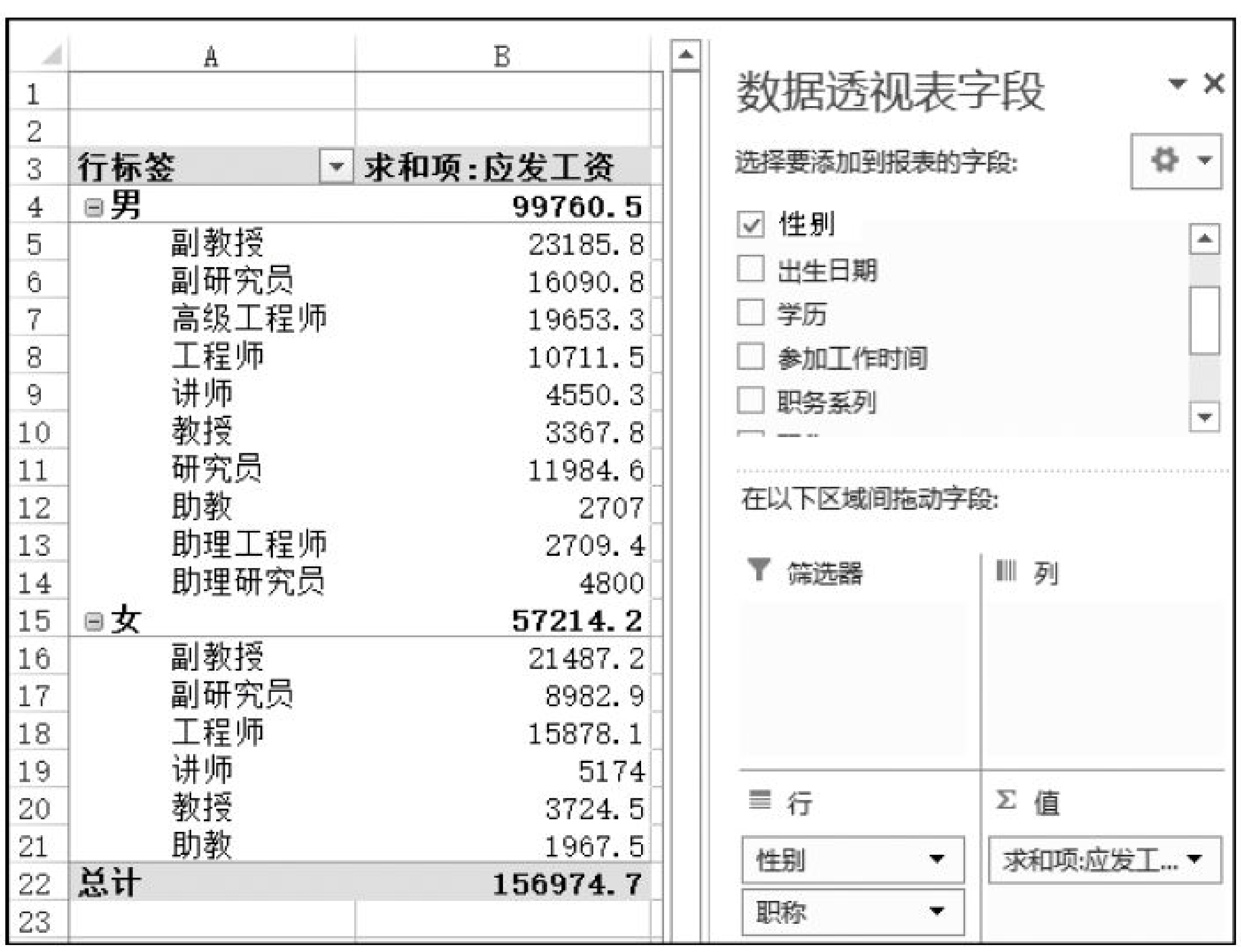


图 4-11 数据透视表按行显示

(3) 若在“数据透视表字段” 窗格中将字段“性别”选项拖动到“列”处,数据可按“行”“列”分别显示,这种透视表的效果很常用,如图 4-12 所示。

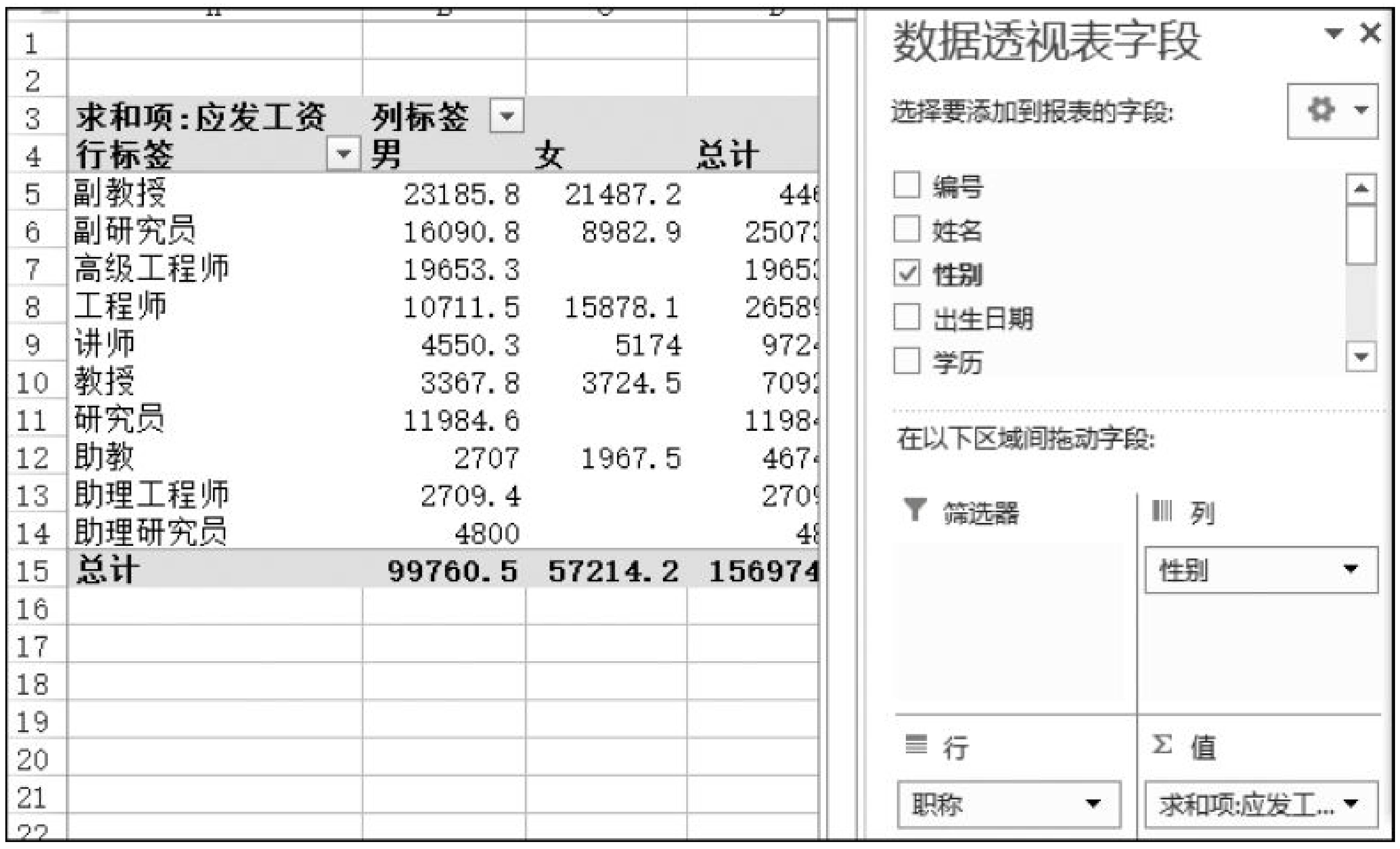


图 4-12　数据透视表按行、列分别显示

第三节　统 计 分 组

数据经过预处理后，可进一步进行分组整理。统计分组是数据整理中的一个重要工作，它是根据统计研究的需要，将数据按照某种特征或标准分成不同的组别。分组时所依据的特征或标准称为统计分组标志。

一、统计分组的含义

统计分组是根据统计研究的目的和任务要求，按照统计分组标志将总体划分成性质不同的若干个部分或组别，使组和组之间具有差异性，而同一组内具有同质性。

二、统计分组的作用

统计分组的作用有以下几点。

（一）区分事物的性质

如企业按照经济性质分组，可分为国有经济、集体经济、私营经济、个体经济、外商投资经济、港澳台经济。

（二）研究事物内部结构

如将国民生产总值按照三次产业划分，可计算出各个产业所占比重，以便研究内部结构是否合理。

（三）研究现象之间的关系

在统计分组的基础上，研究现象和现象之间的相互依存关系。如施肥量和亩产量之

间的关系;商业企业规模和商品流通费用率之间的关系等。

三、统计分组的种类

统计分组的种类有以下两种。

(一) 按分组标志的性质不同,分为品质标志分组和数量标志分组

品质标志分组进一步分为分类标志分组和顺序标志分组。标志表现比较简单,很容易确定组与组之间的性质界限。

数量标志分组又进一步分为单变量值分组和组距分组两种。

单变量值分组是将一个变量值作为一组,适合于离散变量以及变量值较少的情况。

组距分组是将变量值的一个区间作为一组。

组距分组适合于连续变量,适合于变量值较多的情况,可采用等距分组,也可采用不等距分组。等距分组各组频数的分布不受组距大小的影响,可直接根据绝对频数来观察频数分布的特征。不等距分组各组频数的分布受组距大小不同的影响,各组绝对频数的多少不能反映频数分布的实际状况,需要用频数密度(频数密度=频数/组距)反映频数分布的实际状况。在组距分组中,一个组的最小值称为下限,一个组的最大值称为上限。

采用组距式分组时,要遵循"不重不漏"的原则。"不重"是指一项数据只能分在其中的某一组,不能在其他组中重复出现;"不漏"是指组别能够穷尽所有分类,不能有所遗漏。为解决"不重"的问题,统计分组时可以按照"上限不在内"的原则,即当相邻两组的上下限重叠时,恰好等于某一组上限的变量不算在本组内,而计算在下一组内。

在组距分组中,有上限值和下限值的分组称为闭口组;但如果全部数据中的最大值和最小值与其他数据相差悬殊,为了避免出现空白组(没有变量值的组)或个别极端值被漏掉,第一组和最后一组可以采取"××以下"及"××以上"这样的开口组。

组距式分组掩盖了各组内的数据分布状况,为反映各组数据的一般水平,我们通常用组中值作为该组数据的一个代表值。组中值的具体计算方法为

闭口组的组中值=(下限+上限)÷2

缺少下限开口组的组中值=上限-邻组组距÷2

缺少上限开口组的组中值=下限+邻组组距÷2

利用组中值来代表各组变量值的一般水平实际上是以组内各单位的变量值呈均匀分布或以组中值为中心呈对称分布为前提的,但实际上大多数资料并非如此。因此,用组中值作为各组变量值的代表值有误差。

(二) 按分组标志的多少不同,分为简单分组和复合分组

简单分组是指对总体按照一个标志进行的分组。如对某一班级的同学们按性别分组,对某一班级的同学们按学习成绩分组就是两次简单分组。

总体按照两个及两个以上标志结合起来(层叠或交叉)进行的分组就是复合分组。如对某一班级的同学们按性别分组,再对某一班级的同学们按学习成绩层叠或交叉分组就是一次复合分组。交叉式的复合分组得到的信息量最大,能够充分利用调查资料。

第四节　分 布 数 列

按某种标志对数据进行分组后，再计算出所有类别或数据在各组中出现的次数或频数，就形成了一张频数分布表。我们把全部数据按其分组标志在各组中的分布状况称为频数分布或次数分布，又称分布数列。分布在各组内的数据个数称为频数或次数，各组频数与全部频数之和的比值称为频率。对数据进行分组的过程就是频数分布形成的过程。

一、品质型数据的频数分布

品质型数据包括分类数据和顺序数据，它们在分组整理上大多是相同的，但也存在一些小的差异。

（一）分类数据的频数分布

分类数据本身就是对事物的分类，因此在分组时每一种类别就是一组，我们只需计算出每一类的频数、频率或百分比、比率等，即可形成分类型数据的频数分布表。

1. 频数、频率与频数分布

频数是指落在各类别中的数据个数；频率又称比例，是指某一类别数据占全部数据的比值；把各个类别及落在其中的频数和对应频率全部列出，并用表格的形式表示出来，称为频数分布。标准的分类数据频数分布表的格式如表 4-1 所示。

表 4-1　××频数分布表

类　别	频　数	频率（比例）
类别 1 类别 2 ⋮ 类别 n		
合　计		

2. 百分比、比率

对于分类数据，可以用百分比、比率等统计量进行描述性分析。百分比是指将比例乘以 100%得到的结果，是将对比的基数抽象化为 100 而计算出来的。

比率是各种不同类别数值之间的比值。在对经济和社会问题的研究中经常使用比率，如国内生产总值中第一、第二、第三产业产值之比等。

（二）顺序数据的频数分布

上面介绍的分类数据的分组与频数分布的相关概念，如频数、频率、百分比、比率等，也都适用于对顺序数据的整理。但一些适用于顺序数据的整理方法，并不适用于分类数据。对于顺序数据，除了可使用分类数据的整理方法外，还可以计算累积频数和累积频率。

（1）累积频数。累积频数是将各有序类别或组的频数逐级累加起来得到的频数。频

数的累积方法有两种：一种是向上累积，即从类别顺序的开始一方向类别顺序的最后一方累加频数；一种是向下累积，即从类别顺序的最后一方向类别顺序的开始一方累加频数。通过累积频数，可以很容易看出某一类别以下或某一类别以上的频数之和。

(2) 累积频率。累积频率是将各有序类别或组的频率逐级累加起来的频率，它也有向上累积和向下累积两种方法。

标准的顺序数据频数分布表的格式如表 4-2 所示。

表 4-2　××频数分布表

类别	频数	频率（百分比）	向上累积		向下累积	
			频数	频率	频数	频率
类别 1						
类别 2						
⋮						
类别 n						
合　计			—	—	—	—

(三) Excel 在品质型数据频数分布表制作中的应用

在 Excel 中，计算品质型数据的频数可用 COUNTIF() 函数，COUNTIF() 函数用来返回计算区域中满足给定条件的单元格的个数。其语法为

```
COUNTIF(range,criteria)
```

其中，range 为需要计算其中满足条件的单元格数目的单元格区域；criteria 为确定哪些单元格被计算在内的条件，其形式可以为数字、表达式和文本。例如，条件可以表示为"32""<32"">32"或"苹果"等。

下面举例说明在 Excel 2013 中制作品质型数据频数分布表的方法。

【例 4-5】 对图 4-1 中的人员职称结构进行统计，试制作出该校员工的职称数据的分组与频数分布表。

在 Excel 中的操作步骤如下。

(1) 新建一张工作表，把表 4-1 中的职称数据复制到新建表的第一列。

(2) 录入频数分布表的标题数据，如图 4-13 所示。



	A	B	C	D	E	F	G
1	职称		员工职称频数分布表				
2	副教授		职称	人数	百分比	向上累积频率	向下累积频率
3	副教授		教授				
4	高级工程师		副教授				
5	高级工程师		研究员				
6	副研究员		副研究员				
7	研究员		高级工程师				
8	副教授		工程师				
9	研究员		合计				
10	副教授						

图 4-13　员工职称频数分布表(1)

(3) 录入频数数据计算公式。在 D3 单元格中录入公式：=COUNTIF($A$2:$A$19,C3)，或者单击“公式”→“插入函数”，选择“全部函数”类别，找到 COUNTIF 函数，单击“确定”按钮，在弹出的对话框中，在 range 内输入“$A$2:$A$19”，在 criteria 中输入“C3”，单击“确定”按钮。

(4) 复制公式。选中 D3 单元格，将光标移到单元格的右下角，当光标变为“**+**”形状时按住左键不放，向下拖曳至 D8 单元格时，松开左键，效果如图 4-14 所示。

	A	B	C	D	E	F
1	职称		员工职称频数分布表			
2	副教授		职称	人数	百分比	向上累积频率
3	副教授		教授	1		
4	高级工程师		副教授	4		
5	高级工程师		研究员	2		
6	副研究员		副研究员	4		
7	研究员		高级工程师	3		
8	副教授		工程师	4		
9	研究员		合计			
10	副教授					

图 4-14 员工职称频数分布表(2)

(5) 录入人数合计公式。在 C9 单元格录入公式：“=SUM(D3:D8)”，按 Enter 键。

(6) 录入百分比计算公式。在 E3 单元格录入公式“=D3/$D$9”，按 Enter 键。

(7) 复制公式。选中 E3 单元格，将光标移到单元格的右下角，当光标变为“**+**”形状时按住左键不放，向下拖曳至 D9 单元格时，松开左键。

(8) 录入向上累积频率公式。在 F3 单元格内输入“=E3”，按 Enter 键，然后在 F4 单元格内输入“=E4+F3”，按 Enter 键。

(9) 复制公式。选中 F4 单元格，将光标移到单元格的右下角，当光标变为“**+**”形状时按住左键不放，向下拖曳至 F8 单元格时，松开左键。

(10) 录入向下累积频率公式。在 G3 单元格内输入“=E9”，按 Enter 键，然后在 G4 单元格内输入“=G3-E3”，按 Enter 键。

(11) 复制公式。选中 G4 单元格，将光标移到单元格的右下角，当光标变为“**+**”形状时按住左键不放，向下拖曳至 G8 单元格时，松开左键。

(12) 设置格式。把表格格式设置成表 4-2 的标准格式(步骤略)，最终结果如图 4-15 所示。

	A	B	C	D	E	F	G
1	职称		**员工职称频数分布表**				
2	副教授		**职称**	**人数**	**百分比**	**向上累积频率**	**向下累积频率**
3	副教授		教授	1	6%	6%	100%
4	高级工程师		副教授	4	22%	28%	94%
5	高级工程师		研究员	2	11%	39%	72%
6	副研究员		副研究员	4	22%	61%	61%
7	研究员		高级工程师	3	17%	78%	39%
8	副教授		工程师	4	22%	100%	22%
9	研究员		**合计**	18	100%	-	-
10	副教授						

图 4-15 员工职称频数分布表(3)

二、数值型数据的频数分布

上面介绍的分类和顺序数据的整理方法,也完全适用于数值型数据的分组整理,但数值型数据还有特定的整理方法,它们并不适用于分类数据和顺序数据。

(一)单项式变量分布数列的编制

编制单项式变量值频数分布的步骤如下。

(1)对原始数据排序。

(2)确定分组组数,每个单变量值为一组,从原始数据中找到单变量值的个数。

(3)对各组数据所出现的频数进行计数。

在 Excel 中,计算单变量值分组的频数也可用 COUNTIF()函数,下面举例说明在 Excel 2013 中编制单项式变量值频数分布表的步骤。

【例 4-6】 某生产车间 50 名工人日加工零件数(单位:个)如下所示。试编制单变量值频数分布数列。

117	122	124	129	139	107	117	130	122	125
108	131	125	117	122	133	126	122	118	108
110	118	123	126	133	134	127	123	118	112
112	134	127	123	119	113	120	123	127	135
137	114	120	128	124	115	139	128	124	121

Excel 2013 中的操作步骤如下。

(1)将原始数据放到 A1:A50,并进行排序。

(2)输入频数分布表标题及单项式变量值分组数据,如图 4-16 所示。

	A	B	C	D	E	F	G	H
1	107							
2	108		日加工零件数	工人数	百分比	日加工零件数	工人数	百分比
3	108		107			123		
4	110		108			124		
5	112		110			125		
6	112		112			126		
7	113		113			128		
8	114		114			127		
9	115		115			129		
10	117		117			130		
11	117		118			131		
12	117		119			133		
13	118		120			134		
14	118		121			135		
15	118		122			137		
16	119					139		
17	120		合计					

图 4-16 单变量值频数分布表(1)

(3)输入频数计算公式,在 D3 单元格输入公式:"=COUNTIF($A$1:$A$50,C3)"。

(4)复制公式。选中 D3 单元格,将光标移到单元格的右下角,当光标变为"**+**"形状时按住左键不放,向下拖曳至 D15 单元格时,松开左键。复制 D3 单元格,粘贴至 G3 单元格,选中 G3 单元格,将光标移到单元格的右下角,当光标变为"**+**"形状时按住左键不放,向下拖曳至 G16 单元格时,松开左键。

（5）输入合计公式。在 D17 单元格中输入公式：“＝SUM(D3:D15,G3:G16)”，按 Enter 键。

（6）输入百分比计算公式。在 E3 单元格内输入公式：“＝D3/$D$17”。

（7）复制公式。

（8）设置成标准频数分布表格式，如图 4-17 所示。

	A	B	C	D	E	F	G	H
1	107							
2	108		日加工零件数	工人数	百分比	日加工零件数	工人数	百分比
3	108		107	1	2%	123	4	8%
4	110		108	2	4%	124	3	6%
5	112		110	1	2%	125	2	4%
6	112		112	2	4%	126	2	4%
7	113		113	1	2%	128	2	4%
8	114		114	1	2%	127	3	6%
9	115		115	1	2%	129	1	2%
10	117		117	3	6%	130	1	2%
11	117		118	3	6%	131	1	2%
12	117		119	1	2%	133	2	4%
13	118		120	2	4%	134	2	4%
14	118		121	1	2%	135	1	2%
15	118		122	4	8%	137	1	2%
16	119					139	2	4%
17	120		合计	50		–		

图 4-17 单变量值频数分布表(2)

（二）组距式变量分配数列的编制

编制步骤如下。

（1）对数据进行排序，或通过粘贴函数 max 和 min，找出最大值和最小值，计算出全距。$R=x_{max}-x_{min}$。

（2）确定分组组数。实际分组时，可参考美国学者斯特杰斯(H. A. Sturges)创用的经验公式：

$$K=1+3.3\lg N$$

式中，K 为组数，N 为总体单位数或数据的个数；对结果用四舍五入法取整数，即为组数。当然，这只是个经验公式，具体应用时，还要考虑数据的多少、特点和统计分析的要求。

（3）确定组距和组限。组距是一个组的上限与下限之差。组距也可借助 Sturges 的经验公式来确定：

$$d=\frac{R}{K}=\frac{x_{max}-x_{min}}{1+3.3\lg N}$$

式中，d 为组距，R 为全距，即最大变量值 x_{max} 与最小变量值 x_{min} 之差。

最小组下限要小于等于最小变量值，最高组上限要大于等于最大变量值。习惯上，组距和组限确定为 5 的倍数。

（4）对各组数据所出现的频数进行计数。

Excel 中的频数分布函数能对数值型数据进行分组，建立频数分布，较好地描述数据分布状态。Excel 的频数分布函数为 FREQUENCY()函数，FREQUENCY()函数以一列垂直数组返回某个区域中数据的频率分布。

频数分布函数语法形式为

```
FREQUENCY(data_array,bins_array)
```

其中,data_array 为一数组或对一组数值的引用,用来编制频数分布的数据;bins_array 为间隔的数组或对间隔的引用,该间隔用于对 data_array 中的数值进行分组,即为频数或次数的接受区间。

下面举例说明在 Excel 2013 中编制组距分组频数分布表的步骤。

【例 4-7】 利用例 4-6 中的数据编制组距分组频数分布数列。

Excel 中的操作步骤如下。

(1) 将原始数据或排序后的数据放到 A1:A50,把各组的上限(注意:上组限不在内原则)确定好(各组上限依次为:109、114、119、124、129、134、139)放到 C6:C12,如图 4-18 所示。

(2) 单击选择存放频数(工人人数)的单元格 D6:D12。

(3) 单击粘贴函数找到 FREQUENCY,或输入 =FREQUENCY(),在弹出的对话框分别输入 A1:A50 和 C6:C12,使用 Ctrl+Shift+Enter 组合键,即得到频数分布,如图 4-18 所示。

	A	B	C	D
1	107	139	107	32
2	108	组数		7
3	108	组距		5
4	110			
5	112	日加工零件分组		工人人数
6	112	105~	109	3
7	113	110~	114	5
8	114	115~	119	8
9	115	120~	124	14
10	117	125~	129	10
11	117	130—	134	6
12	117	135~	139	4
13	118	合	计	50
14	118			

图 4-18 组距分组

第五节 统计汇总技术

一、统计汇总形式

统计汇总形式有逐级汇总、集中汇总和综合汇总三种。

(一) 逐级汇总

逐级汇总是自下而上地逐级汇总调查资料。优点是能满足各级对资料的需要;缺点是差错大,费时。

(二) 集中汇总

集中汇总又称超级汇总,就是将原始资料集中到最高一级机关汇总,减少了汇总环节,节约了汇总时间,减少了汇总差错。

(三) 综合汇总

综合汇总就是将逐级汇总和集中汇总两种形式结合。一方面对基本的统计指标实施逐级汇总,另一方面对全部资料实行集中汇总。如我国人口普查就采用了这种汇总方式,这样既保证了必要人口资料的及时使用,又保证了对人口情况进行深入分析的需要。要花费较多的人力、物力和财力。

二、统计汇总技术

统计汇总技术有手工汇总和电子计算机汇总两种。

（一）手工汇总

手工汇总有画记号法、过录法、折叠法和卡片法。

1. **画记号法**

画记号法就是用点、线组成笔画是 5 或 10 的字或图形来计算各组单位数的方法，最常用的是画“正”字（5 画）法。

2. **过录法**

过录法就是将需要汇总的调查资料先过录到事先设计好的汇总表上，然后计算出各组单位数的方法。

3. **折叠法**

折叠法就是将需要汇总的调查资料、需要汇总的项目或数据折在边或角上，将其叠放在一起进行汇总的方法。

4. **卡片法**

卡片法就是将需要汇总的调查资料先过录到特制的卡片上，一个调查单位的资料录在一张卡片上，利用卡片进行归类汇总的方法。

（二）电子计算机汇总

电子计算机汇总一般要经过编制程序、编码、数据录入、逻辑检查和制表打印几个步骤。应用电子计算机进行统计资料的汇总，其主要特点是速度快、精度高，在程序控制下自动工作，进行各种逻辑判断，并能存储大量的数据信息。

第六节　统　计　表

统计表和统计图是显示统计数据的两种方式。在日常生活中，阅读报纸杂志，或者在看电视、查阅计算机网络时，都能看到大量的统计表格和统计图形。统计表把杂乱的数据有条理地组织在一张简明的表格内，统计图把数据形象地显示出来。显然，看统计表和统计图要比看一堆枯燥的数字更有趣。当我们自己对某些实际问题进行研究时，也经常要使用统计表和统计图。正确地使用统计表和统计图是做好统计分析的最基本技能。

一、统计表的构成

统计表是用于显示统计数据的基本工具。在数据的收集、整理、描述和分析过程中，都要使用统计表。许多杂乱的数据，既不便于阅读，也不便于理解和分析，一旦整理在一张统计表内，就会使这些数据变得一目了然、清晰易懂。充分利用和绘制好统计表是做好统计分析的基本要求。

统计表的形式多种多样，根据使用者的要求和统计数据本身的特点，可以绘制形式多

样的统计表。比如,表 4-3 就是一种比较常见的统计表。

表 4-3 2005—2015 年我国国内生产总值数据 } 表头

亿元

列标题 / 行标题 / 数字资料

年 份	国内生产总值	其 中		
		第一产业	第二产业	第三产业
2005	183 867.9	23 070.4	87 364.6	73 432.9
2006	210 871.0	247 37.0	103 162.0	82 972.0
2007	270 232.3	115 810.7	126 633.6	27 788.0
2008	319 515.5	136 805.8	149 956.6	32 753.2
2009	349 081.4	154 747.9	160 171.7	34 161.8
2010	413 030.3	182 038.0	191 629.8	39 362.6
2011	489 300.6	216 098.6	227 038.8	46 163.1
2012	540 367.4	244 821.9	244 643.3	50 902.3
2013	595 244.4	277 959.3	261 956.1	55 329.1
2014	643 974.0	308 058.6	277 571.8	58 343.5
2015	676 708.0	60 863.0	274 278.0	341 567.0

资料来源:国家统计局. 中国统计年鉴 2016[M]. 北京:中国统计出版社,2016. } 附加

从表 4-3 可以看出,统计表一般由四个主要部分组成,即表头、行标题、列标题和数字资料,必要时可以在统计表的下方加上附加。表头应放在表的上方,它所说明的是统计表的主要内容。行标题和列标题通常安排在统计表的第一行和第一列,它所表示的主要是所研究问题的类别名称和指标名称,通常也被称为“类”。如果是时间序列数据,行标题和列标题也可以是时间,当数据较多时,通常将时间放在行标题的位置。表的其余部分是具体的数字资料。附加通常放在统计表的下方,主要包括资料来源、指标的注释和必要的说明等内容。

二、统计表的设计

由于使用者的目的以及统计数据的特点不同,统计表的设计在形式和结构上会有较大差异,但设计上的基本要求则是一致的。总体上看,统计表的设计应符合科学、实用、简练、美观的要求。具体来说,设计统计表时要注意以下几点。

(1) 要合理安排统计表的结构,如行标题、列标题、数字资料的位置应安排合理。当然,由于强调的问题不同,行标题和列标题可以互换,但应使统计表的横竖长度比例适当,避免出现过高或过长的表格形式。

(2) 表头一般应包括表号、总标题和表中数据的单位等内容。

(3) 总标题应简明确切地概括出统计表的内容,一般需要表明统计数据的时间(when)、地点(where)以及何种数据(what),即标题内容应满足 3w 要求。

(4) 如果表中的全部数据都是同一计量单位,可放在表的右上角标明,若各指标的计量单位不同,则应放在每个指标后打括号()表明或单列计量单位栏次。

(5) 表中的上下两条横线一般用反线,中间的其他线要用正线。这样使人看起来清

楚、醒目。通常情况下,统计表的左右两边不封口,列标题之间一般用竖线隔开,而行标题之间通常不必用横线隔开。总之,表中尽量少用横竖线。

(6) 表中的数据一般是右对齐,有小数点时应以小数点对齐,而且小数点的位数应统一。对于没有数字的表格单元,一般用“—”表示,一张填好的统计表不应出现空白单元格。

(7) 对于行合计,既可以放在第一行,也可以放在最后一行,放在第一行后面可只列出其中一部分;对于列合计只能放在最后一列。

(8) 在使用统计表时,必要时可在表的下方加上注释,特别要注意注明资料来源,以表示对他人劳动成果的尊重,方便读者查阅使用。

第七节 统 计 图

统计图是统计资料整理结果的另一种常用的表达方式,它是利用几何图形(点、线、面、形)或其他图形把所研究对象的特征、内部结构等相互关联的数量关系绘制成的简明的图形。统计图表示的数量关系形象、直观、明白,可以使人们一目了然地认识客观事物的状态、形成、发展趋势或在某地区上的分布状况等,故它在经济管理工作中使用非常广泛。

Excel 具有丰富多彩的制图功能,它可以将表中的数据用图形表示,使表、图、文字有机地结合起来。它提供的图形种类繁多,如条形图、饼图、圆环图、折线图、直方图等 100 多种基本图表类型。不同的统计数据需要借助不同的统计图来显示。针对数据类型,我们把种类繁多的统计图形做以下概括,如图 4-19 所示。

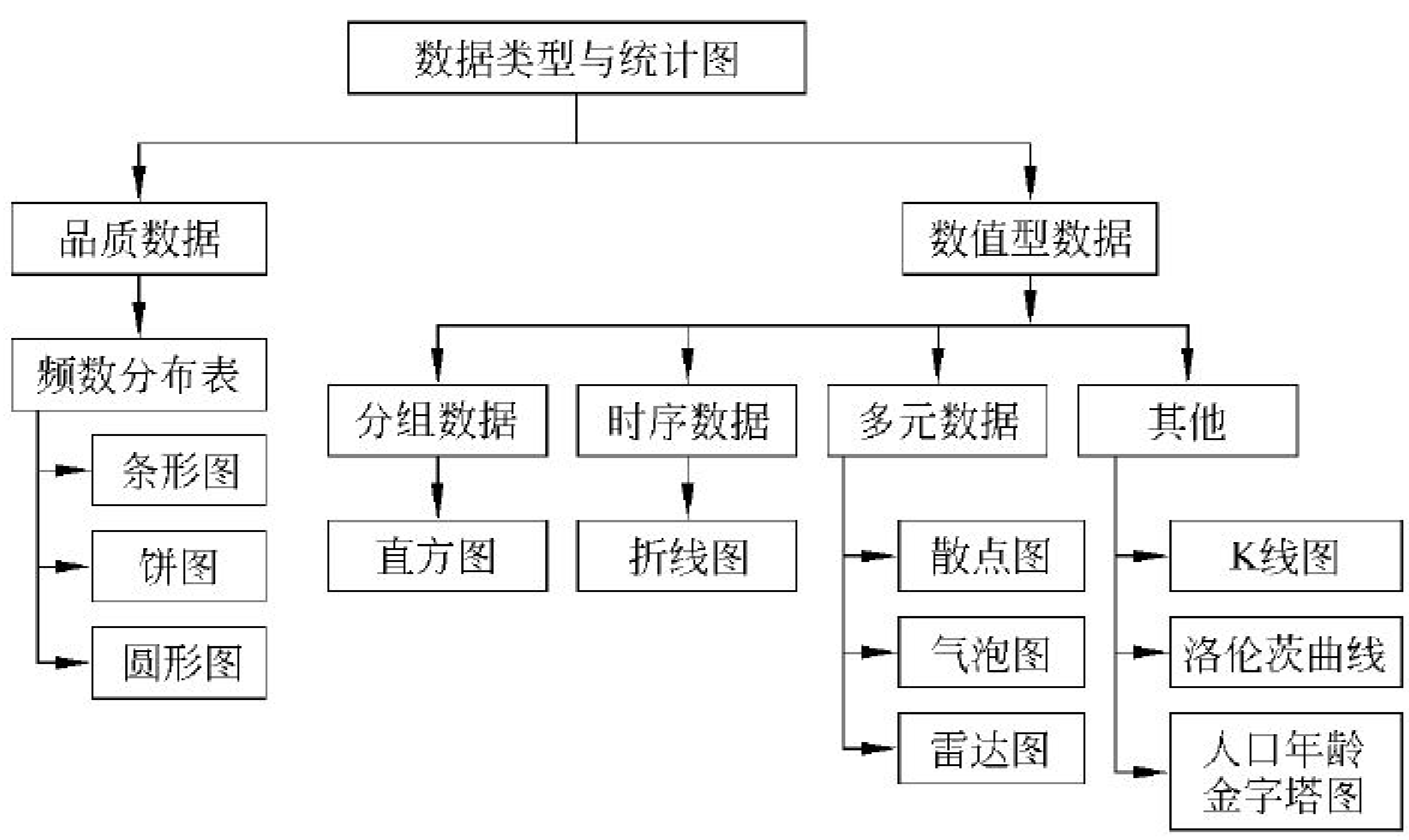


图 4-19 数据类型与统计图

一、反映品质数据的统计图——条形图、饼图和圆环图

(一) 条形图(柱形图)

条形图是用宽度相同的条形的高度或长短来表示各类别数据的图形,有单式条形图、复式条形图等形式,主要用于反映分类数据的频数分布。Excel 的条形图是水平条形图,

柱形图为竖直条形图,绘制时,各类别可以放在纵轴,称为条形图,也可以放在横轴,称为柱形图。两种图都可以用来表示一组或几组分类相关的数值,它可用于不同现象的比较,也可用于同一现象不同时间的比较,在条形图或柱形图中,各条或柱的宽度、各条或柱间的距离彼此均等,条的长度或柱的高度与代表的变量值成比例。

下面举例说明条形图在 Excel 2013 中的制作方法。

【例 4-8】 以图 4-15 中的数据为例,在 Excel 2013 中制作条形图的操作步骤如下。

(1) 打开工作表,用鼠标选定数据区域。

(2) 选择"插入"菜单,单击"推荐图表"→"所有图表",在弹出的窗口中选择"柱形图"选项,如图 4-20 所示。



图 4-20　柱形图(1)

(3) 单击"确定"按钮即可,如图 4-21 所示。

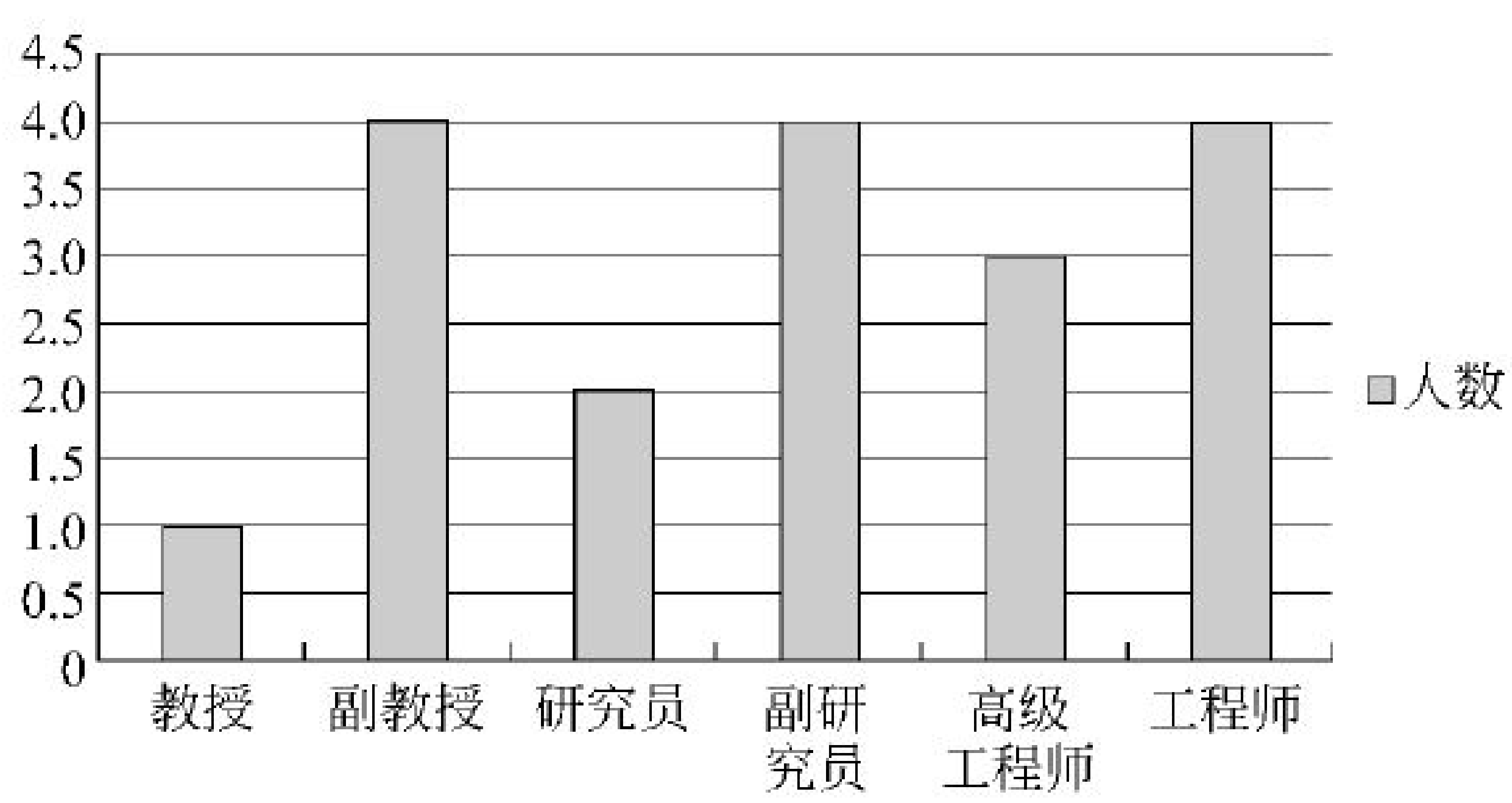


图 4-21　柱形图(2)

(二) 饼图(圆形图)

饼图也称圆形图,是用圆形及圆内扇形的角度来表示数值大小的图形。主要用于表示总体或样本中各组成部分所占的比例,对于研究结构性问题十分有用。绘制圆形图时,总体中各部分所占的百分比用圆内的各个扇形角度表示,这些扇形的中心角度,是按各部

分数据百分比占360°的相应比例确定的。

下面举例说明饼图在 Excel 2013 中的制作方法。

【例 4-9】 以图 4-15 中的数据为例，在 Excel 2013 中制作饼图的操作步骤如下。

（1）打开工作表，用鼠标选定数据区域。

（2）选择“插入”菜单，单击“推荐图表”→“所有图表”，在弹出的窗口中选择“饼图”。

（3）单击“确定”按钮即可，如图 4-22 所示。

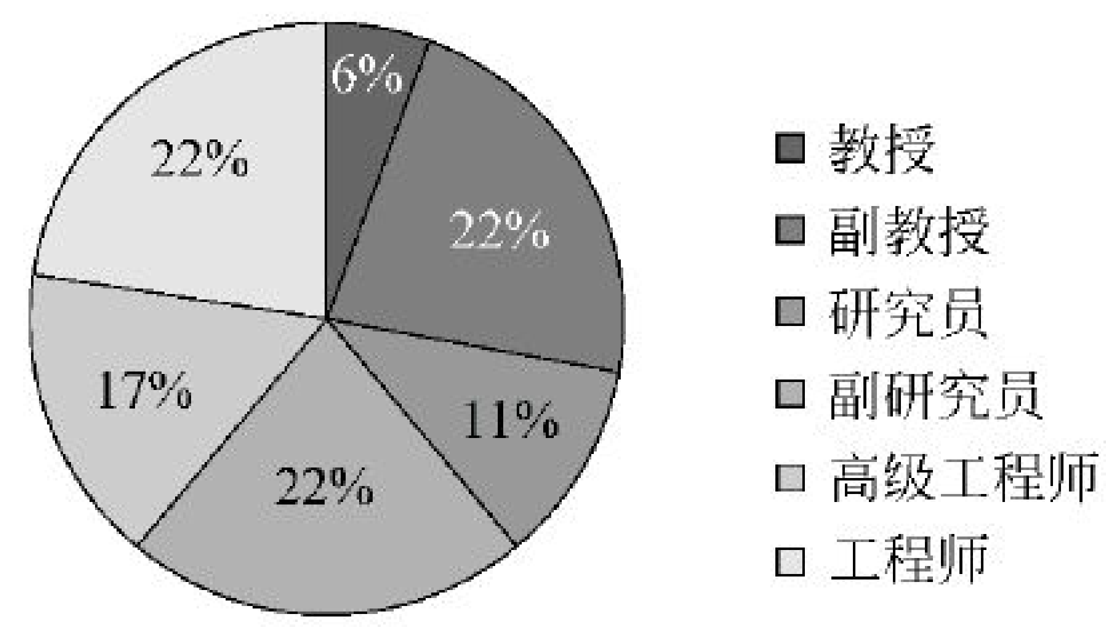


图 4-22　饼图

（三）圆环图

圆环图中间有一个“空洞”，总体中的每一部分数据用环中的一段表示。圆环图与饼图类似，但又有区别：饼图只能显示一个总体各部分所占的比例，圆环图则可以同时绘制多个总体的数据系列，每一个总体的数据系列为一个环。圆环图可用于结构比较研究，主要用于展示分类和顺序数据。

下面举例说明圆环图在 Excel 2013 中的制作方法。

【例 4-10】 在甲、乙两个城市对居民的住房状况满意度进行调查后的频数分布表如表 4-4、表 4-5 所示，根据两个表在 Excel 2013 中制作环形图的操作步骤如下。

表 4-4　甲城市家庭对住房状况评价的频数分布

回答类别	甲城市					
	户数/户	百分比/%	向上累积		向下累积	
			户数/户	百分比/%	户数/户	百分比/%
非常不满意	24	8	24	8	300	100
不满意	108	36	132	44	276	92
一般	93	31	225	75	168	56
满意	45	15	270	90	75	25
非常满意	30	10	300	100	30	10
合　计	300	100	—	—	—	—

表 4-5　乙城市家庭对住房状况评价的频数分布

回答类别	乙城市					
	户数/户	百分比/%	向上累积		向下累积	
			户数/户	百分比/%	户数/户	百分比/%
非常不满意	21	7	21	7	300	100
不满意	99	33	120	40	279	93
一般	78	26	198	66	180	60
满意	64	21.3	262	87.3	102	34
非常满意	38	12.7	300	100	38	12.7
合　计	300	100	—	—	—	—

(1) 打开工作表,把回答类别和甲、乙两个城市的对应频率数据复制到另外的数据区域。

(2) 用鼠标选定数据区域。

(3) 选择“插入”菜单,单击“推荐图表”→“所有图表”,在弹出的窗口中选择“圆环图”。

(4) 单击“确定”按钮即可,如图 4-23 所示。

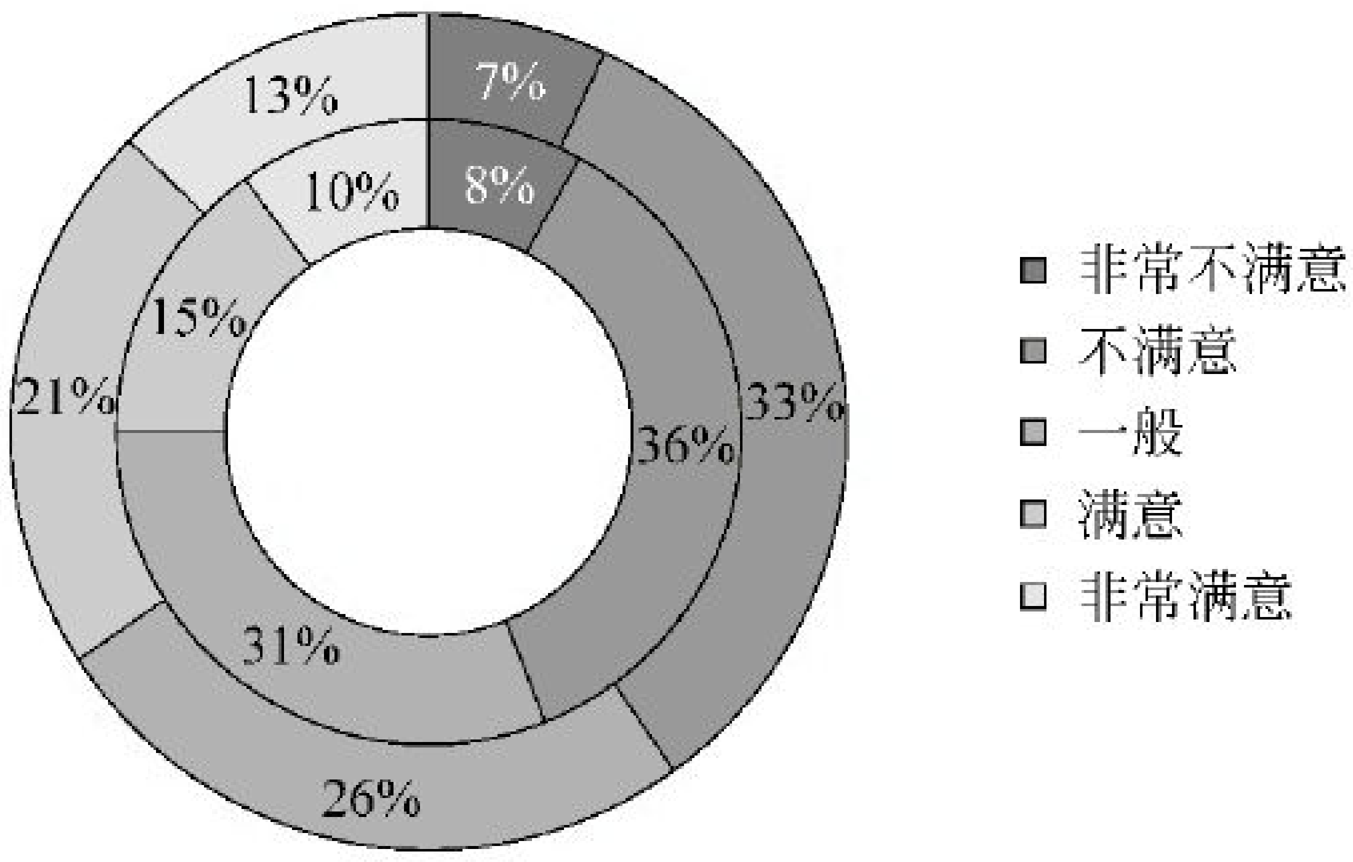


图 4-23 圆环图

二、反映分组数据的统计图——直方图

直方图是用矩形的宽度和高度来表示频数分布的图形,实际上是用矩形的面积来表示各组的频数分布,在直角坐标中,用横轴表示数据分组,纵轴表示频数或频率,各组与相应的频数就形成了一个矩形,即直方图,直方图下的总面积等于 1。

直方图与条形图的区别如下。

(1) 条形图是用条形的长度(横置时)表示各类别频数的多少,其宽度(表示类别)则是固定的。

(2) 直方图是用面积表示各组频数的多少,矩形的高度表示每一组的频数或百分比,宽度则表示各组的组距,其高度与宽度均有意义。

(3) 直方图的各矩形通常是连续排列,条形图则是分开排列。

(4) 条形图主要用于展示分类数据,直方图则主要用于展示数值型数据。

利用 Excel 2013 绘制直方图要用到“数据分析”工具。(第一次使用见第一章第五节。)

下面举例说明在 Excel 中制作直方图的方法。

【例 4-11】 以图 4-16 中的数据为例,在 Excel 2013 中制作直方图的操作步骤如下。

(1) 单击“数据”→“数据分析”,此时弹出“数据分析”对话框。如图 4-24 所示。

(2) 在弹出的“数据分析”对话框中选择“直方图”选项,再单击“确定”按钮,进入直方图对话框,如图 4-25 所示。

(3) 在“直方图”对话框中输入、单击相应的项目(图 4-25),再单击“确定”按钮,就得到如图 4-26 所示的结果。



图 4-24 直方图(1)



图 4-25 直方图(2)

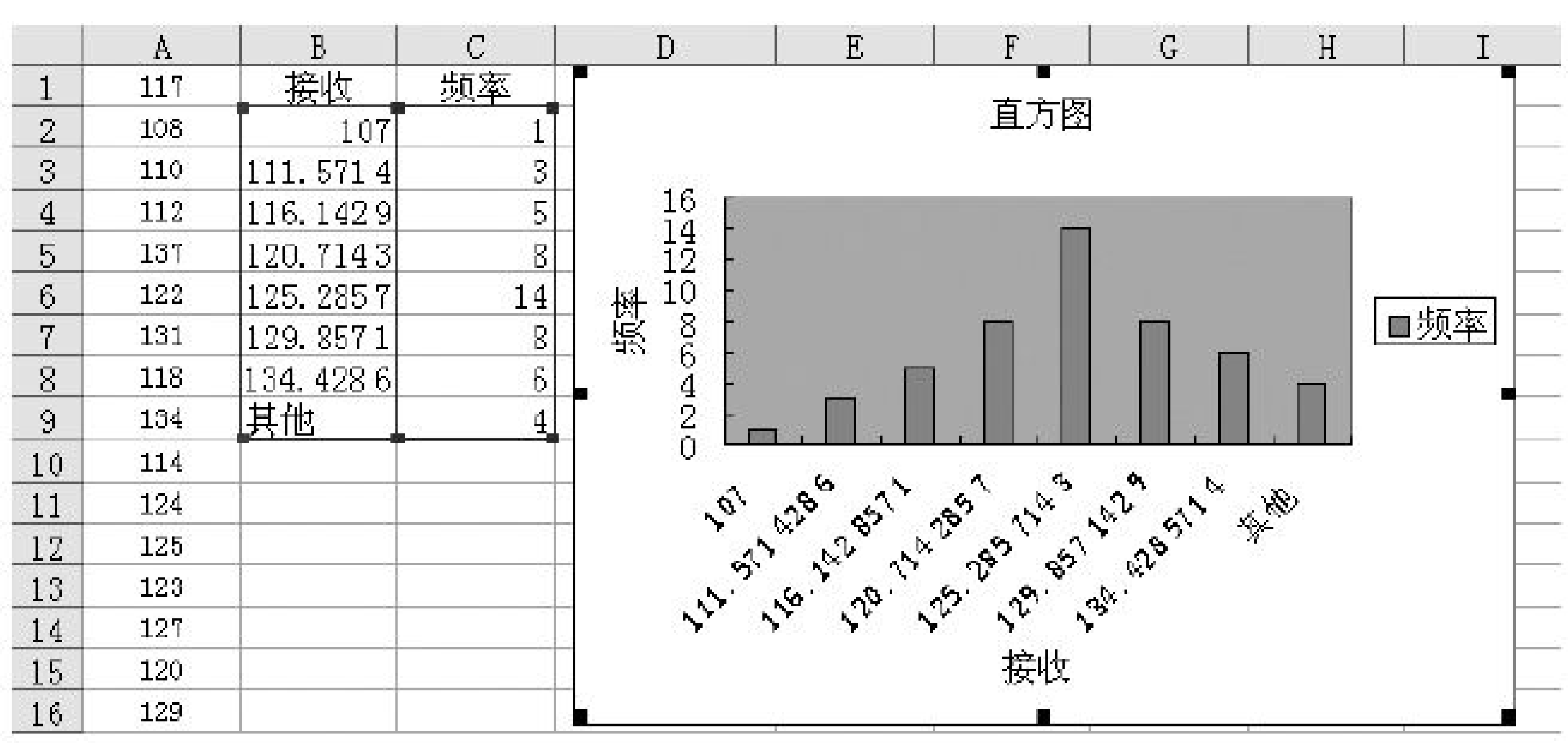


图 4-26 直方图(3)

(4) 修改图 4-26。修改的办法是：选中图 4-26 中的 B2:B8 单元格，右击，选择“设置单元格格式”选项，在弹出的单元格格式中选择数值，将小数位数调至 0，单击“确定”按钮，就得到如图 4-26 所示的结果。

在此需要说明的是直方图对话框中的接收区域：在此输入接收区域的单元格引用，该区域应包含一组可选的用来定义接受区域的边界值，这些值应当按升序排列，即统计分组的上限。如果省略此处的接受区域，Excel 将在数据组的最小值和最大值之间创建一组平滑分布的接受区间(输出结果如图 4-26 所示)。

(5) 修改图 4-27。为了使图更容易理解，需要将各个柱形连接起来，可对图 4-27 做如下修改。

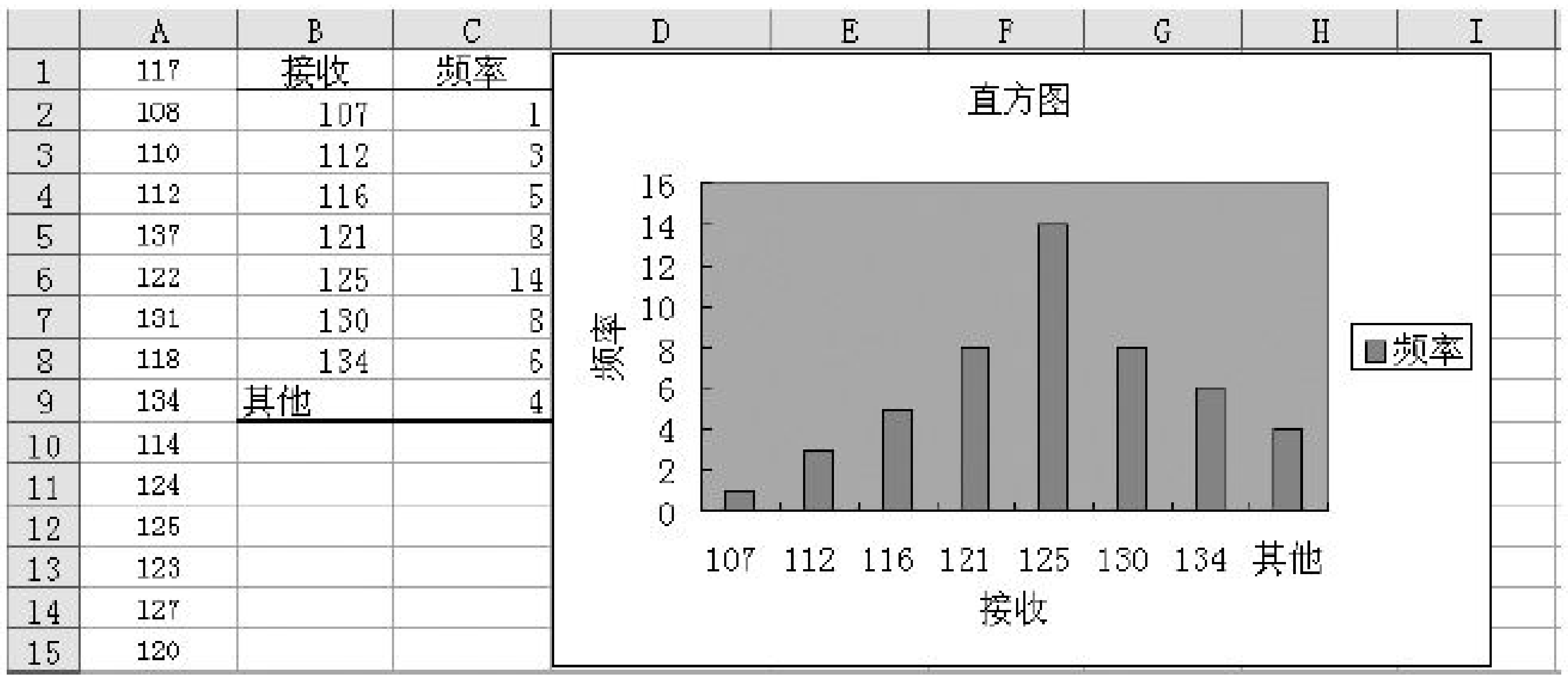


图 4-27　直方图(4)

选中图 4-27 中的某个柱形,右击,在弹出的菜单中,选择设置“数据系列格式”格式,此时弹出“数据系列格式”对话框,在弹出的“数据系列格式”对话框左侧,单击系列选项,将右侧分类间距调整为“无间距”,单击“确定”按钮,即得到如图 4-28 所示的结果。

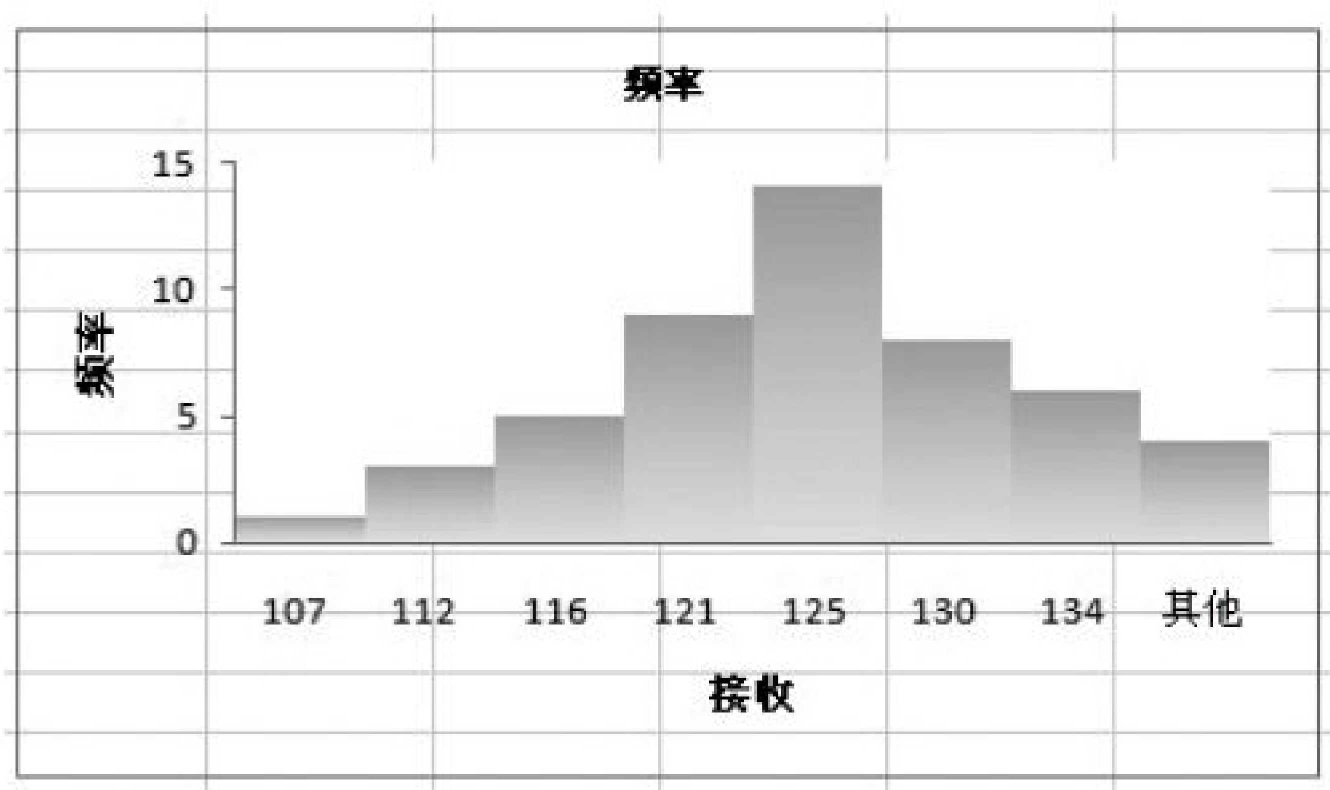


图 4-28　直方图(5)

三、反映时序数据的统计图——折线图

时序数据,即时间序列数据。它由两部分数据组成:一是时间数据,二是对应时间的观测值(变量值)。绘制线图(line plot)时应注意以下几点:①时间一般绘在横轴,指标数据绘在纵轴;②图形的长宽比例要适当,其长宽比例大致为 10∶7;③一般情况下,纵轴数据下端应从“0”开始,以便于比较。数据与“0”之间的间距过大时,可以采取折断的符号将纵轴折断。

下面举例说明折线图在 Excel 2013 中的制作方法。

【例 4-12】 已知 2005—2014 年我国城乡居民家庭的人均收入数据如表 4-6 所示。试绘制折线图。

表 4-6　2005—2014 年城乡居民家庭人均收入　　元

▲	A	B	C
1	年份	乡村居民	城镇居民
2	2005	2 784	9 832
3	2006	3 066	10 739
4	2007	3 538	12 480

续表

▲	A	B	C
5	2008	4 065	14 061
6	2009	4 402	15 127
7	2010	4 941	17 104
8	2011	6 187	19 912
9	2012	6 964	21 861
10	2013	7 773	23 609
11	2014	8 711	25 424

绘制的折线图如图 4-29 所示。

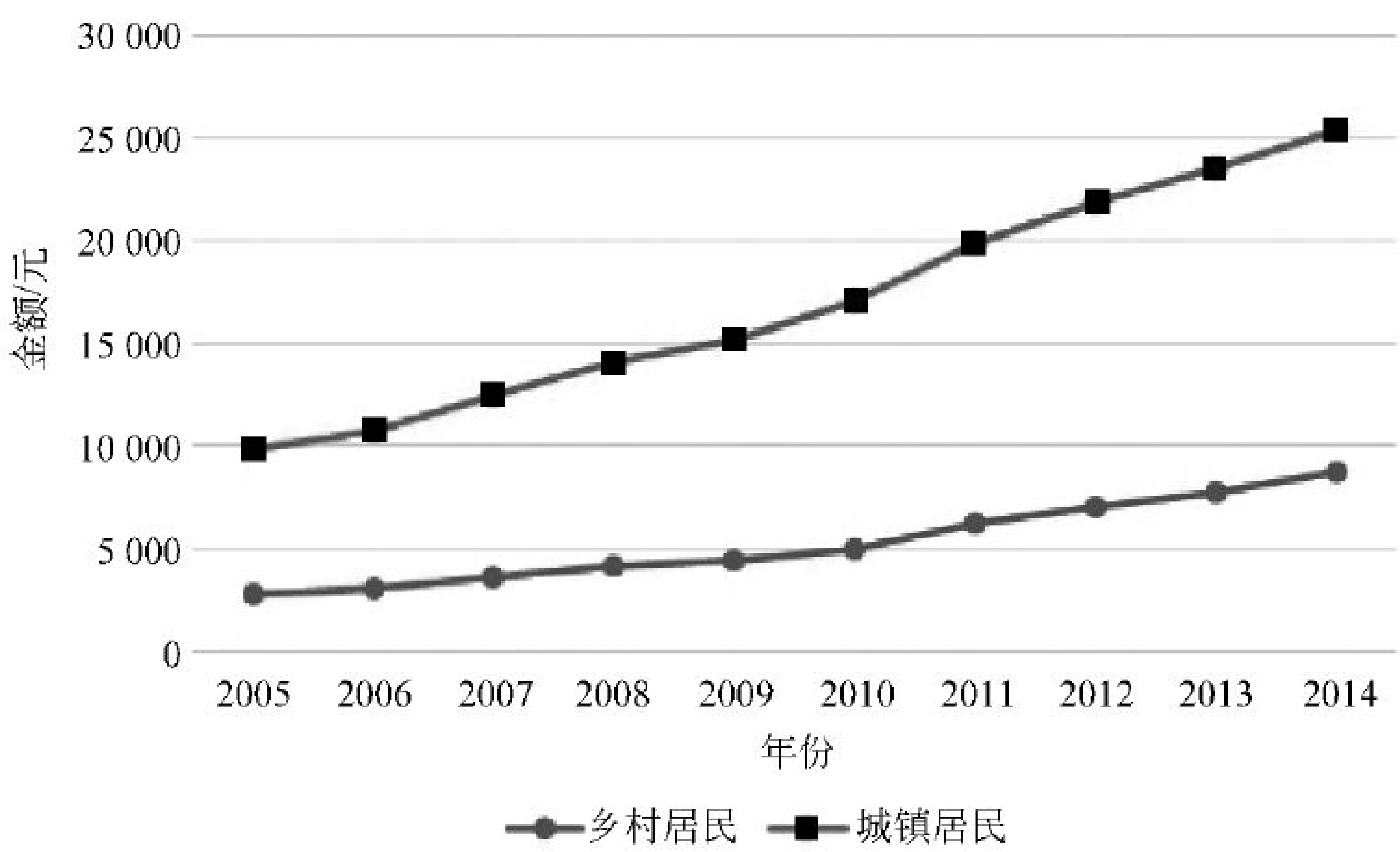


图 4-29 2005—2014 年城乡居民家庭人均收入折线图

在 Excel 2013 中的操作步骤如下。

(1) 打开工作表,输入数据,如表 4-6 所示。

(2) 用鼠标选定人均收入的数据区域,在本例中为 B1:C11。

(3) 选择“插入”菜单,单击“推荐图表”→“所有图表”,在弹出的窗口中选择“折线图”。

(4) 单击“确定”按钮,如图 4-29 所示。

四、反映多元数据的统计图——散点图、气泡图、雷达图

(一) 散点图

散点图是用二维坐标展示两个变量之间关系的一种图形,可用于观察两个变量之间的一种数量上的变化关系。

下面举例说明在 Excel 中的制作方法。

【例 4-13】 某产品的温度与其密度和黏度有一定的关系。为了了解它们之间的关系,收集到相关数据,如图 4-30 所示。试绘制该产品温度和黏度的散点图,并分析它们之

间的关系。

Excel 2013 中的操作步骤如下。

(1) 用鼠标选定数据区域,在本例中为 B3:C11,选择“插入”菜单,选择“推荐图表”→“所有图表”,在弹出的窗口中选择“XY(散点图)”以及子图:仅带数据标记的散点图。

(2) 单击“确定”按钮,添加上行标题和列标题即可得到散点图,如图 4-31 所示。

	A	B	C
1	某产品温度与密度和黏度的数据		
2	密度	黏度	温度
3	0.457	3.55	500
4	0.525	3.25	400
5	0.616	2.93	300
6	0.675	2.75	250
7	0.746	2.57	20
8	0.835	2.38	150
9	0.946	2.17	100
10	1.09	1.95	50
11	1.29	1.71	0

图 4-30 某产品温度与黏度的数据

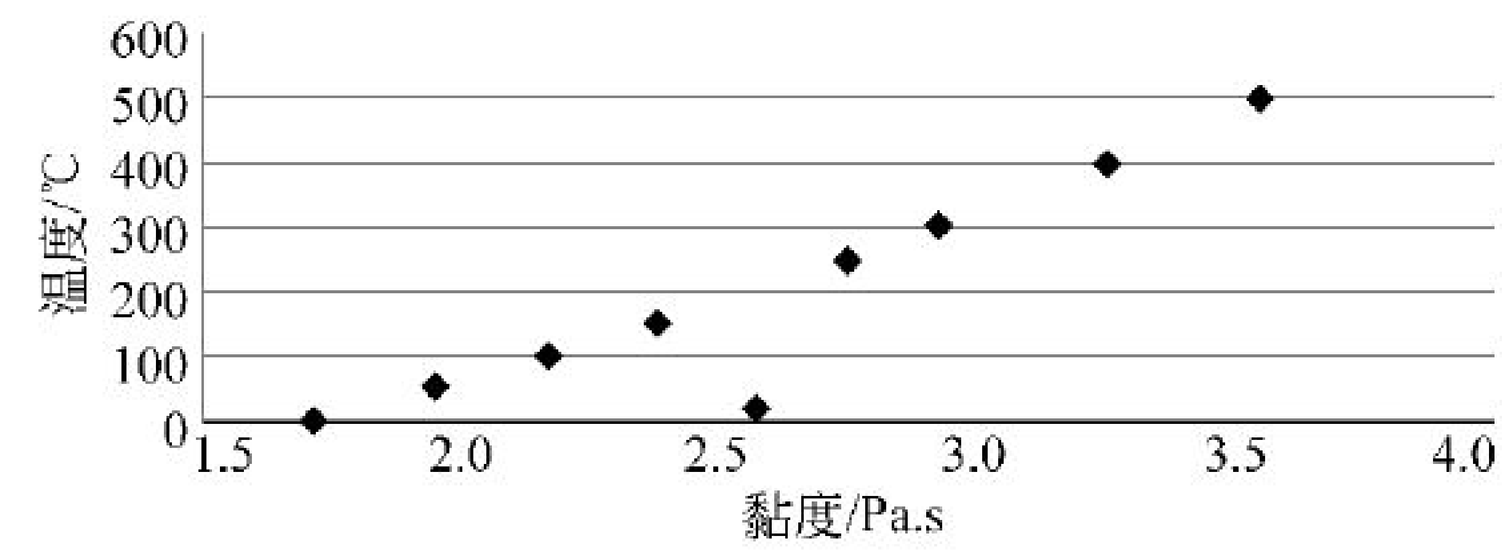


图 4-31 散点图

从图 4-31 可以看出,该产品的温度与黏度之间具有明显的线性关系,随着温度的升高,黏度也逐渐变大。

(二) 气泡图

气泡图是用于展示三个变量之间的关系的一种图形。它与散点图的显示原理类似,用横轴和纵轴分别表示一个变量,而第三个变量用气泡的大小来表示。

例如,根据图 4-30 所列数据绘制的气泡图如图 4-32 所示。

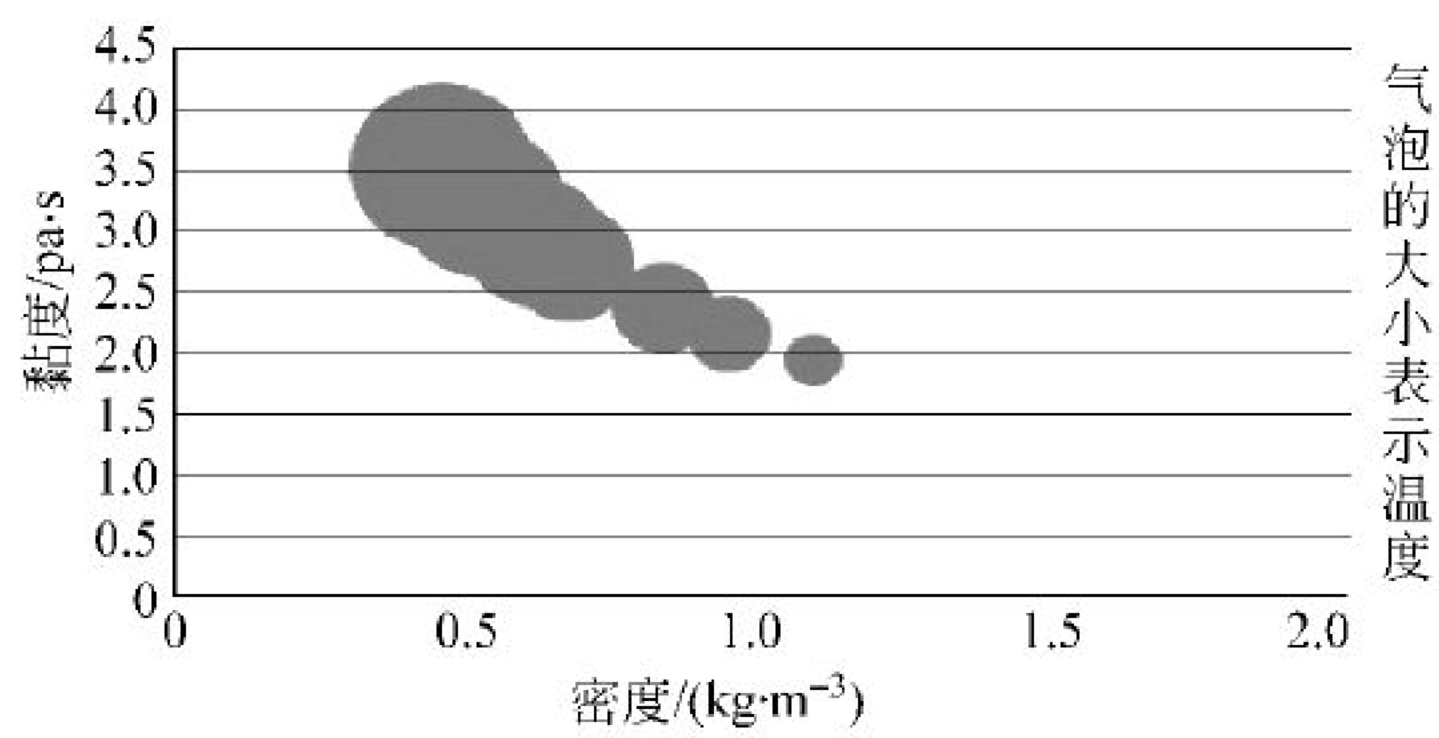


图 4-32 气泡图

在 Excel 2013 中绘制气泡图的方法类似于散点图,不再赘述。

(三) 雷达图

雷达图是显示多个变量的图示方法,在显示或对比各变量的数值总和时十分有用。假定各变量的取值具有相同的正负号,总的绝对值与图形所围成的区域成正比。可用于研究多个样本之间的相似程度。

设有 n 组样本 $S_1, S_2, \cdots, S_n$,每个样本测得 P 个变量 $X_1, X_2, \cdots, X_P$,要绘制这 P 个变量的雷达图,其具体做法如下。

（1）做一个圆，然后将这个圆 P 等分，得到 P 个点，令这 P 个点分别对应 P 个变量；再将这 P 个点与圆心连线，得到 P 个辐射状的半径，这 P 个半径分别作为 P 个变量的坐标轴，每个变量值的大小由半径上的点到圆心的距离表示。

（2）将同一样本的值在 P 个坐标上的点连线。这样，n 个样本形成的 n 个多边形就是一个雷达图。

下面举例说明雷达图在 Excel 中的制作方法。

【例 4-14】 2016 年某地区城乡居民家庭平均每人各项生活消费支出构成数据如表 4-7 所示，试绘制雷达图。

表 4-7　2016 年某地区城乡居民家庭平均每人生活消费支出构成　　%

项　　目	城镇居民	农村居民
食品	36.52	33.69
衣着	10.47	6.62
家庭设备用品及服务	10.02	22.04
医疗保健	6.42	5.83
交通通信	6.98	11.5
娱乐教育文化服务	13.72	9.72
居住	12.01	8.2
杂项商品与服务	3.87	2.4

在 Excel 2013 中的操作步骤如下。

（1）打开工作表，输入数据，如图 4-33 所示。

（2）用鼠标选定数据区域，在本例中为 A1:C9。

（3）选择"插入"菜单，单击"其他图表"→"所有图表类型"，在弹出的窗口中选择"雷达图"选项，如图 4-34 所示。

（4）单击"确定"按钮，如图 4-35 所示。



	A	B	C
1	项　目	城镇居民	农村居民
2	食品	36.52	33.69
3	衣着	10.47	6.62
4	家庭设备用品及服务	10.02	22.04
5	医疗保健	6.42	5.83
6	交通通信	6.98	11.5
7	娱乐教育文化服务	13.72	9.72
8	居住	12.01	8.2
9	杂项商品与服务	3.87	2.4

图 4-33　录入数据



图 4-34　雷达图(1)

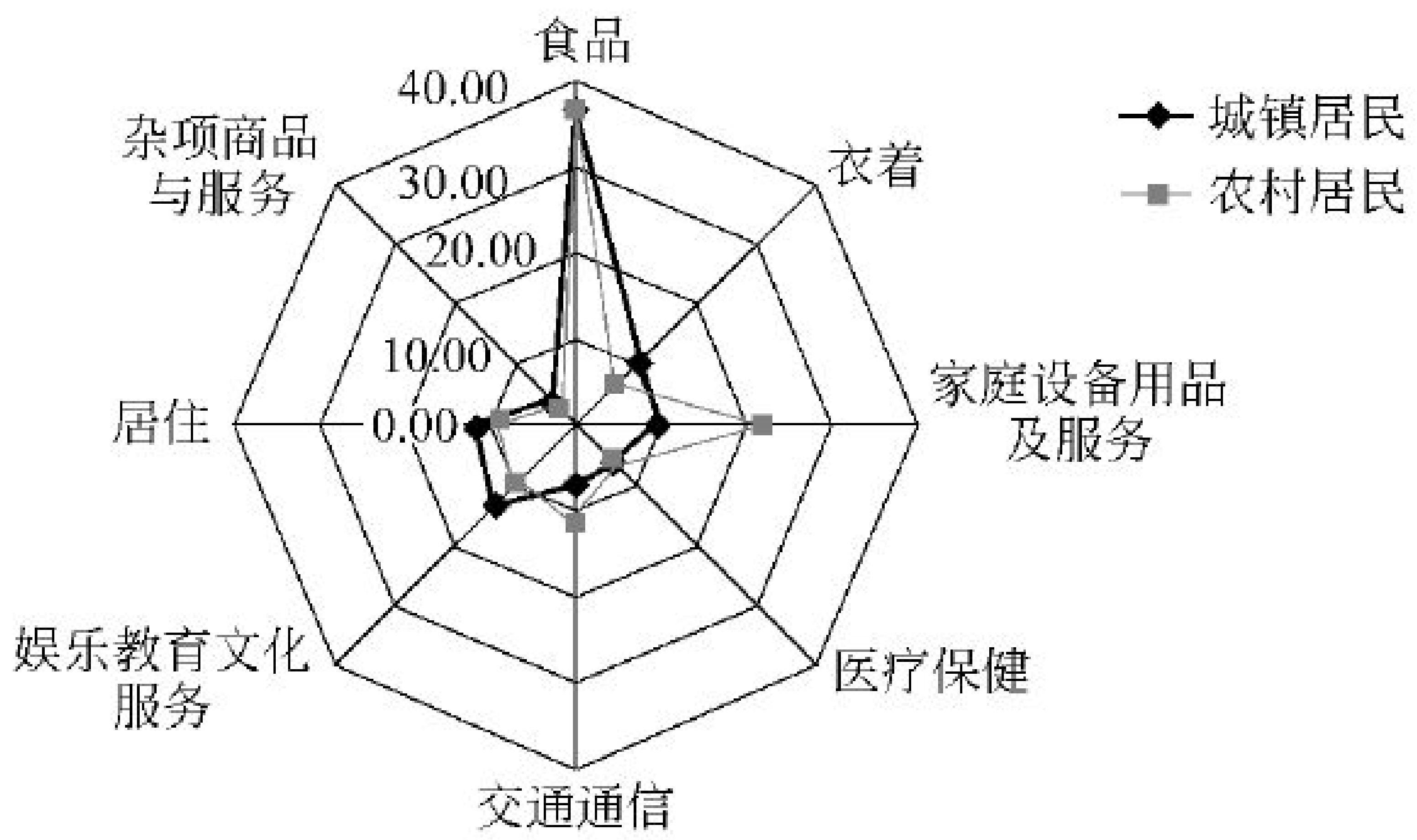


图 4-35　雷达图(2)

思　考　题

1. 数据的预处理包括哪些内容?
2. 品质型数据如何进行分组?
3. 数值型数据如何进行分组?
4. 统计表制作时应注意的问题有哪些?
5. 直方图与条形图有哪些区别?
6. 饼图与圆环图有哪些区别?

练　习　题

1. 为了调查某地区手机品牌的市场占有率状况,随机抽取了 50 名人员进行调查,调查结果如表 4-8 所示。

表 4-8　某地区手机品牌的市场占有率状况

三星	三星	三星	摩托罗拉	联想
诺基亚	其他	三星	联想	诺基亚
索尼爱立信	诺基亚	诺基亚	其他	诺基亚
三星	三星	摩托罗拉	联想	联想
长虹	索尼爱立信	三星	索尼爱立信	摩托罗拉
诺基亚	联想	诺基亚	三星	其他
三星	摩托罗拉	其他	摩托罗拉	诺基亚
诺基亚	其他	诺基亚	长虹	诺基亚
联想	诺基亚	诺基亚	其他	三星
其他	其他	诺基亚	诺基亚	其他

要求:

(1) 根据表 4-8 的数据,用 Excel 制作一张频数分布表。

(2) 绘制一张柱形图,反映其分布。

(3) 绘制一张圆形图反映其分布。

2. 某调查公司为了获取消费者对现有手机的满意情况,分别从 A、B 两个地区随机抽取 500 名消费者进行了调查,调查结果如表 4-9 和表 4-10 所示。

表 4-9　A 城市消费者对现有手机的满意度情况

回答类别	A 城市					
	户数/户	百分比/%	向上累积		向下累积	
			户数/户	百分比/%	户数/户	百分比/%
非常不满意	34					
不满意	258					
一般	143					
满意	105					
非常满意	60					
合　计	600		—	—	—	—

表 4-10　B 城市消费者对现有手机的满意度情况

回答类别	B 城市					
	户数/户	百分比/%	向上累积		向下累积	
			户数/户	百分比/%	户数/户	百分比/%
非常不满意	80					
不满意	110					
一般	178					
满意	144					
非常满意	88					
合　计	600		—	—	—	—

要求:

(1) 在 Excel 中编制表 4-9 和表 4-10,并利用公式把剩余的数据补齐。

(2) 绘制反映两个城市满意度状况的圆环图。

3. 某商场连续 60 天的商品销售额如表 4-11 所示。

表 4-11　某商场连续 60 天的商品销售额　　万元

62	42	71	45	45	39	50	53	54	44
69	54	56	74	68	48	57	56	63	48
53	44	51	51	66	65	54	62	63	60
63	68	56	54	63	39	60	60	65	62
38	69	57	56	57	48	66	59	60	59
54	56	56	59	65	65	66	68	57	63

要求:

(1) 根据表 4-11 中的数据在 Excel 中进行适当的分组。

(2) 编制频数分布表。

(3) 根据分组数据绘制直方图和折线图。

4. 2010 年某手机商城手机销售年度销量排名前 10 的外国品牌和国产品牌销售数据如表 4-12 所示。

表 4-12 品牌手机销售数据 台

外国手机品牌销售排行	销售量	国产手机品牌销售排行	销售量
苹果 iPhone 4	834	联想 Lenovo	373
诺基亚 3720 classic	812	U9 手机 U9mobile	255
三星 i9000	555	多普达 Dopod	170
诺基亚 C5	447	金立 Gionee	75
诺基亚 N8	416	海尔 Haier	75
HTC HD2	300	ZTE 中兴	58
诺基亚 E52	285	夏新 Amoi	43
诺基亚 6700c	272	波导 Bird	40
诺基亚 6303ci	218	康佳 KONKA	25
HTC Desire HD	176	海信 Hisense	15

要求:

(1) 画出外国手机品牌和国产手机品牌销售量的对比条形图。

(2) 画出外国手机品牌和国产手机品牌销售量的圆环图。

5. 为研究某条河流的污染程度,环保局分别在河流的上、中、下游设立了取样点,每个取样点都化验水中的五项污染指标,所得数据如表 4-13 所示。

表 4-13 不同取样点的化验指标数据

河流	指标 1	指标 2	指标 3	指标 4	指标 5
上游	4.52	5	483	196	14
中游	0.34	1.4	36	41	6
下游	2.17	6.8	208	112	35

要求:

(1) 将各指标用雷达图表示出来。

(2) 分析该河流的主要污染源。

6. 某企业所生产的五种产品 6 月的销售数量、销售额及市场占有率情况如表 4-14 所示。

要求:

(1) 在 Excel 中绘制出各种产品的市场占有率的饼图。

(2) 绘制出产品数量、销售额以及市场占有率的气泡图。

表 4-14　五种产品 6 月情况

产品名称	产品数量/件	销售额/件	市场占有率/%
甲产品	14	12 200	12
乙产品	20	60 000	33
丙产品	18	24 400	10
丁产品	22	32 000	42
戊产品	25	46 000	26

7. 已知 2005—2015 年我国国内生产总值数据如表 4-15 所示。

表 4-15　2005—2015 年我国国内生产总值数据　　亿元

年份	国内生产总值	其　中		
		第一产业	第二产业	第三产业
2005	183 867.9	23 070.4	87 364.6	73 432.9
2006	210 871.0	24 737.0	103 162.0	82 972.0
2007	270 232.3	115 810.7	126 633.6	27 788.0
2008	319 515.5	136 805.8	149 956.6	32 753.2
2009	349 081.4	154 747.9	160 171.7	34 161.8
2010	413 030.3	182 038.0	191 629.8	39 362.6
2011	489 300.6	216 098.6	227 038.8	46 163.1
2012	540 367.4	244 821.9	244 643.3	50 902.3
2013	595 244.4	277 959.3	261 956.1	55 329.1
2014	643 974.0	308 058.6	277 571.8	58 343.5
2015	676 708.0	60 863.0	274 278.0	341 567.0

资料来源：国家统计局. 中国统计年鉴 2016[M]. 北京：中国统计出版社，2016.

要求：

(1) 在 Excel 中绘制 2005—2015 年我国国内生产总值及其三次产业的折线图；

(2) 根据 2015 年的国内生产总值及其三次产业的构成数据绘制饼图。

8. 某地区居民的人均医疗保健消费支出明细如表 4-16 所示。

表 4-16　某地区居民人均医疗保健消费支出明细　　元

医疗保健	城镇居民	农村居民
(一) 医疗器具	10	1
(二) 保健器具	23	2
(三) 药品费	382	72
(四) 滋补保健品	135	2
(五) 医疗费	314	139
(六) 其他	21	4

要求:

(1) 根据表 4-16 中的数据,绘制城镇居民和农村居民人均医疗保健消费支出对比条形图和圆环图。

(2) 绘制雷达图,分析城镇居民和农村居民人均医疗保健消费支出的差别。

9. 依据表 4-17 所示数据绘制散点图。

表 4-17 练习题 9 题图

1	2.5	3.2	2.4	1.5	2.7	1.1	0.8	0.9
12	30	57.6	60	31.5	97.2	29.7	26.4	25.2

第五章

平均分析

【统计知识 ABC】

[A]平均工资

职工平均工资指企业、事业、机关单位的职工在一定时期内平均每人所得的货币工资额。它表明一定时期职工工资收入的高低程度，是反映职工工资水平的主要指标。其计算公式为

职工平均工资＝报告期实际支付的全部职工工资总额/报告期全部职工平均人数

职工工资总额指各单位在一定时期内直接支付给本单位全部职工的劳动报酬总额。工资总额的计算原则应以直接支付给职工的全部劳动报酬为根据。各单位支付给职工的劳动报酬以及其他根据有关规定支付的工资，不论是计入成本的还是不计入成本的；不论是按国家规定列入计征奖金税项目的，还是未列入计征奖金税项目的；不论是以货币形式支付的还是以实物形式支付的，均包括在工资总额内。

（资料来源：http://www.stats.gov.cn/tjzd/tjzbjs/t20020327_14287.html.）

联合国国际劳工组织日前对全球 72 个国家和地区的人均月收入做了最新统计。这些国家和地区的人均月收入是 1 480 美元，约合人民币 9 327 元。其中，中国员工的月平均工资为 656 美元，约合人民币 4 134 元，位列所调查 72 个国家和地区的第 57 位。(2012 年 4 月 4 日《京华时报》)

（资料来源：http://news.jinghua.cn/351/c/201204/04/n3674159.html.）

2011 年城镇单位就业人员平均工资全国为 41 799 元，山东省为 37 618 元。

（资料来源：http://www.stats.gov.cn/tjsj/ndsj/2012/indexch.html.）

[B]城镇居民人均可支配收入

城镇居民人均可支配收入是指家庭总收入扣除交纳的个人所得税、个人交纳的各项社会保障支出后按家庭人口平均可用于支配的收入。可支配收入计算公式为

可支配收入＝家庭总收入－交纳的所得税－个人交纳的社会保障支出－记账补贴

家庭总收入指生活在一起的所有家庭成员在调查期得到的工资性收入、经营性收入、财产性收入、转移性收入的总和，不包括出售财物和借贷收入。收入的统计标准以实际发生的数额为准，无论收入是补发还是预发，只要是调查期得到的都应如实计算，不做分摊。

(1) 工资性收入，指就业人员通过各种途径得到的全部劳动报酬，包括所从事主要职业的工资以及从事第二职业、其他兼职和零星劳动得到的其他劳动收入。

(2) 经营性收入，指家庭成员从事生产经营活动所获得的净收入，使全部生产经营收入扣除生产经营成本和税金(不包括个人所得税)后的收入。

(3) 财产性收入，指家庭拥有的动产(如银行存款、有价证券)、不动产(如车辆、土地、收藏品等)所获得的收入。包括出让财产使用权所获得的利息、租金、专利收入；财产营运所获得的红利收入、财产增值收益等。

(4) 转移性收入，指国家、单位、社会团体对居民家庭的各种转移支付和居民家庭间的收入转移。包括政府对个人收入转移的离退休金、失业救济金、赔偿等；单位对个人收入转移的辞退金、保险索赔、住房公积金等；家庭、亲友间的赠送和赡养等。

个人所得税指调查对象被扣缴的工资薪金所得、对企事业单位的承包经营承租经营所得、劳务报酬所得、稿酬所得、特许权使用费所得、利息股息红利所得、财产租赁所得、财产转让所得、偶然所得、经国务院财政部门确定征税的其他所得等个人所得的税款。

社会保障支出指调查户家庭成员参加国家法律、法规规定的社会保障项目中由个人交纳的保障支出。不包括职工所在单位交纳的那部分社会保障金。具体包括个人交纳的住房公积金、个人交纳的医疗基金、个人交纳的失业基金、个人交纳的养老基金和个人交纳的其他社会保障支出。

记账补贴指调查户因承担记账工作从统计部门、工作单位和其他途径所得到的现金。不包括实物部分。

我国城镇居民人均可支配收入 1978 年为 343.4 元，2011 年为 21 809.8 元。

(资料来源：http://baike.baidu.com/view/2332455.html.)

[C]农民人均纯收入

农民纯收入，指农村住户当年从各个来源得到的总收入相应地扣除所发生的费用后的收入总和。纯收入主要用于再生产投入和当年生活消费支出，也可用于储蓄和各种非义务性支出。计算公式为

农民纯收入＝总收入－家庭经营费用支出－税费支出－生产性固定资产折旧－赠送农村内部亲友

农民人均纯收入，指按人口平均的纯收入水平，反映的是一个地区或一个农户农村居民的平均收入水平。农民人均纯收入的计算公式为

农民人均纯收入＝(农村居民家庭总收入－家庭经营费用支出－生产性固定资产折旧－税金和上交承包费用－调查补贴)/农村居民家庭常住人口

(资料来源：http://baike.baidu.com/view/602333.html.)

第一节　平均分析概述

一、平均分析的含义

平均分析是指利用某一时点的平均指标和差异指标相结合对社会经济现象进行分析的一种方法。

平均指标又称平均数，是同质总体内某一数量标志在一定时间、地点条件下达到的一般水平，即总体内各单位参差不齐的标志值的代表值。平均指标的特点是对数量标志在总体单位之间的数值差异抽象化，它并不代表某一总体单位的具体水平，而是一个用来反映总体综合数量特征的典型水平或者代表水平。例如，用职工平均工资代表职工收入的一般水平，用平均亩产代表粮食生产的一般水平等。

从多数社会经济变量数列的分布来看，通常是接近平均数的变量值的次数较多，而远离平均数的变量值的次数较少，也就是，与平均数离差较小的变量值的次数比较多，而离差较大的变量值的次数比较少，形成正、负离差大致相等的状态。整个变量数列以平均数为中心而上下波动，所以，掌握了变量数列的平均数，就可以了解总体分布集中趋势的一般特征，它是变量分布的重要特征值。

但是，平均指标把总体各单位标志值抽象化了，掩盖了各标志值之间的差异，所以在运用平均指标进行平均分析的同时，有必要进一步对被抽象化的各单位标志值的差异程度进行测度，这种测度指标称为差异指标。差异指标反映的是各变量值远离平均指标值的分散程度，反映各变量值的离中趋势。数据的分散程度越大，平均指标的测度值对该组数据的代表性就越差；分散程度越小，其代表性就越好。

因此，平均指标和差异指标是一对相互联系的对应指标，两者是从不同侧面反映同质总体的共同特征。平均指标说明总体各单位标志值的一般水平，反映总体分布的集中趋势；差异指标则说明总体各单位标志值的差异程度，反映总体分布的离中趋势或分散程度，反映平均指标的代表性大小。

在对数据进行统计整理的基础上进行平均分析，可以找到反映数据分布特征的各个代表值，对数据进行概括性的度量，从而比较全面地把握数据的分布特征。

二、平均分析的作用

平均分析是研究社会经济现象常用的分析方法之一，在社会经济统计中应用十分广泛，在统计工作和统计研究中具有十分重要的作用。

(1) 平均分析可以比较同类现象在不同地区、部门或单位之间的本质差异。例如，比较两个企业职工收入状况，仅比较两个企业的工资总额是不够的，因为工资总额还受职工人数等因素影响，而用人均收入则可以比较好地反映两个企业职工收入的差异。

(2) 平均分析可用于比较同一总体在不同时期的动态发展状况。例如，我国改革开放以来农村居民家庭人均纯收入水平逐年提高，生活逐步得到改善。

(3) 平均分析指标可以为统计推断提供重要的特征值。在统计推断中，用样本平均数推断总体平均数、利用样本方差推断总体方差、确定样本容量，都要用到平均指标和差异指标。例如，在农作物产量调查中，用样本平均亩产推断总平均亩产，进而根据总平均亩产推算总产量。

(4) 平均分析中差异指标可以衡量平均指标的代表性高低。平均指标作为总体各单位某一数量标志的代表值，其代表性的高低与总体各单位变量值的差异程度有直接关系：总体的差异指标值越大，平均数的代表性越低；反之，差异指标值越小，平均数代表性越高。

(5) 平均分析可以衡量社会经济活动的稳定性和均衡程度。例如,检查生产计划执行情况时,除了计算计划完成程度外,还要用差异指标分析计划执行过程中的均衡程度,检查生产过程是否存在时紧时松的现象。又如进行产品质量统计检验,经常采用差异指标,如果指标变动程度越小,说明产品质量越稳定;反之,变动程度越大,则产品质量的稳定性越差。

三、平均分析的种类

平均分析从其内容上来看可以分为平均指标分析和差异指标分析。需要说明的是平均指标分析按其反映的时态不同,可以分为静态平均指标分析和动态平均指标分析。静态平均指标分析主要是依据在同一时间范围内总体各单位某一数量标志值的平均指标进行分析,而动态平均指标分析是依据同一空间不同时间某一指标值的平均水平进行分析。本节着重介绍静态平均指标分析。动态平均指标分析将在时间序列分析一章中介绍。

平均指标分析按其指标度量的方法不同,可以分为数值平均分析和位置平均分析。数值平均分析是根据数列中的每一个数值或变量值计算的平均指标进行分析,指标包括算术平均数、调和平均数和几何平均数;位置平均分析是根据某数值在数列中所处的特殊位置而确定的平均指标进行分析,指标包括中位数、众数、分位数等。

差异分析,本章主要介绍的是总体内部的标志值差异分析,是通过计算相关差异指标进行分析,指标包括异众比率、四分位差、全距、平均差、方差和标准差、离散系数等。

从统计数据的类型来看,不同的数据类型有不同的平均分析指标,但需要说明的是适合于低层次的数据平均分析指标同样适用于高层次的数据;反过来,适合于高层次数据的平均分析指标并不适用于低层次的数据。

平均分析的种类及其指标适用的数据类型如图 5-1 所示。

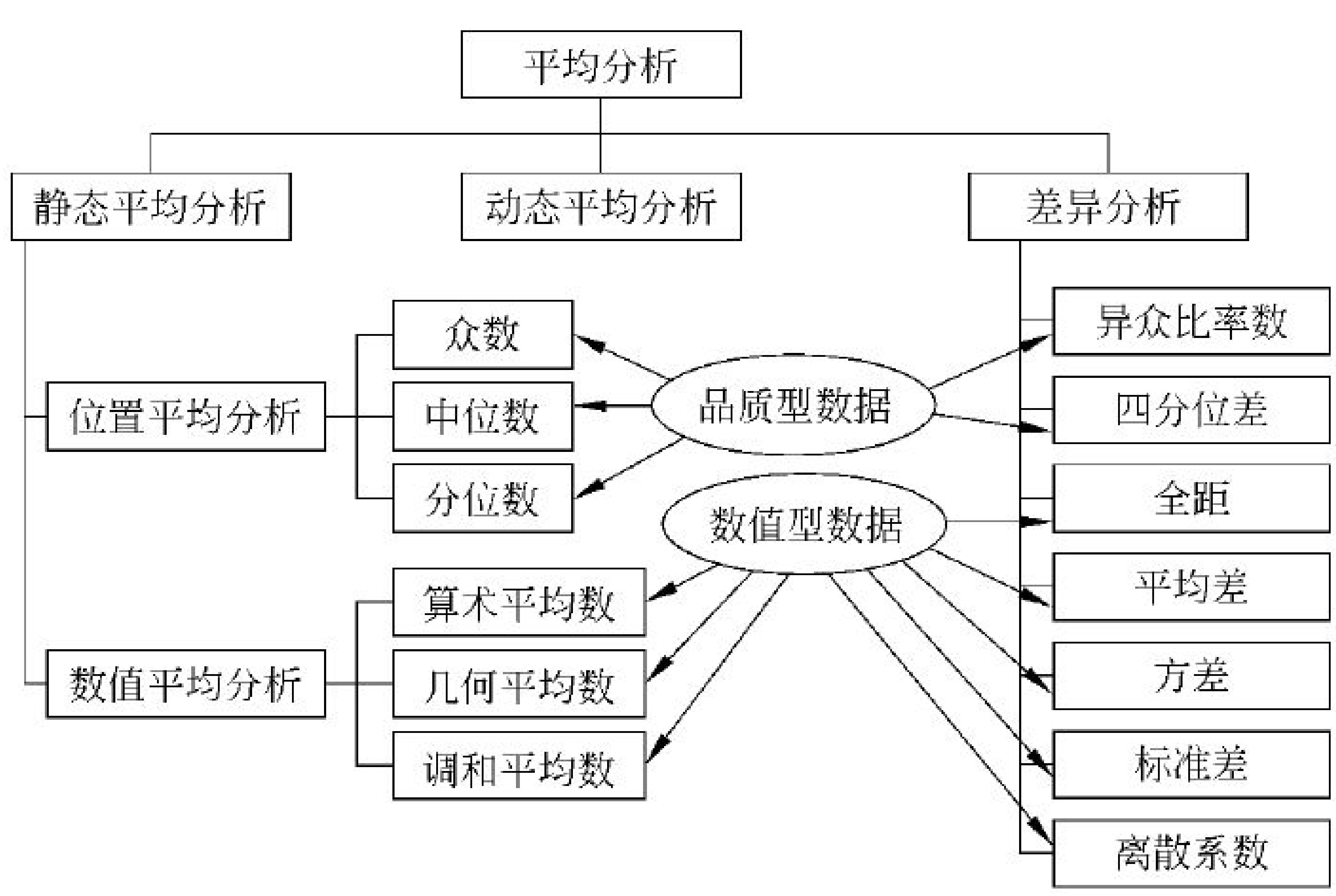


图 5-1 平均分析的种类及其指标适用的数据类型

第二节　静态平均分析

一、众数

众数是指一组数据中出现次数最多的变量值，用 M_o 表示。从变量分布的角度看，众数是具有明显集中趋势点的数值，一组数据分布的最高峰点所对应的数值即为众数。当然，如果数据的分布没有明显的集中趋势或最高峰点，众数也可以不存在；如果有多个高峰点，也就有多个众数。

在 Excel 中计算一组数值中的众数的函数为：MODE()。

其语法结构为：MODE(number1，number2，…)。

其中，number1，number2，…为需要找出众数的 1～30 个参数，也可以是单一数组或者对数据区域的引用。如果没有众数，函数将返回错误值＃N/A。

众数主要用于分类型数据的平均分析，也适用于顺序型数据和数值型数据的平均分析。不同数据类型的分析方法如下。

(1) 对于分类型数据，我们可以首先利用 MAX()函数计算频数数组的最大值，然后找出该数所对应的分组类别就是该分类型数据的众数。

【例 5-1】 根据第四章中图 4-15 中的数据，计算“职称”的众数。

Excel 中具体操作步骤如下。

① 在某个空白单元格中输入函数“＝MAX(D3:D8)”，按 Enter 键，单元格中的计算结果为“4”。

② 找出图 4-14 中人数为 4 的职称“副教授”“副研究员”“工程师”即为所求的众数。

(2) 对于顺序型数据，众数的分析方法等同于分类型数据。

【例 5-2】 根据第四章表 4-4 和表 4-5 中的数据，计算甲、乙两个城市对住房状况满意度的众数。

在 Excel 中的操作步骤如下。

① 把甲、乙两个城市回答类别的频数数据录入表格，如图 5-2 所示。

	A	B	C	D	E	F
1	甲城市	24	108	93	45	30
2	乙城市	21	99	78	64	38

图 5-2　顺序型数据的众数

② 在某个空白单元格中输入函数“＝MAX(B1:F1)”，按 Enter 键，单元格中的计算结果为甲城市回答类别频数的最大值，等于 108。同样计算乙城市回答类别频数的最大值为 99。

③ 找出表 4-4 中频数为 108 的回答类别为“不满意”；表 4-5 中频数为 99 的回答类别为“不满意”。

由此可以看出，两个城市对于住房状况表示不满意的用户最多。

(3) 对于未分组数据，我们可以直接用 MODE()函数，计算出其众数。

【例 5-3】 集贸市场上，某种农作物的成交价格(单位：元/千克)为 1.50，1.60，1.20，1.30，1.40，1.40，1.50，1.70，1.60，1.50，1.50，1.10，1.80，1.90，试计算成交价格的众数。

在 Excel 中的操作步骤如下。

① 把价格数据录入表格,如图 5-3 所示。

	A	B	C	D	E	F	G
1	1.5	1.6	1.2	1.3	1.4	1.4	1.5
2	1.7	1.6	1.5	1.5	1.1	1.8	1.9

图 5-3　未分组数据的众数

② 在某个空白单元格中输入函数"=MODE(A1:G2)",按 Enter 键,单元格中的计算结果为 1.5,即成交价格的众数为 1.5。

由此看出,众数是一个位置平均数,不受极端值的影响,具有稳健性。一般情况下,只有在数据量较大的情况下,次数分布才有明显的集中趋势,众数才有意义。

(4) 对于单变量值分组数据,其众数的计算等同于分类型数据和顺序型数据众数的计算,在此不再赘述。

(5) 对于组距分组数据,我们可以首先找出众数所在的分组,即频数最大的值所在的分组,然后计算众数的具体数值,由于众数的数值与其相邻两组的频数分布有一定的关系,其众数可借助如下计算公式计算。

下限公式:

$$M_o = L + \frac{f_m - f_{-1}}{(f_m - f_{-1}) + (f_m - f_{+1})} \cdot d = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot d \tag{5-1}$$

上限公式:

$$M_o = U - \frac{f_m - f_{+1}}{(f_m - f_{-1}) + (f_m - f_{+1})} \cdot d = U - \frac{\Delta_2}{\Delta_1 + \Delta_2} \cdot d \tag{5-2}$$

式中,L 为众数所在组的下限;U 为众数所在组的上限;f_m 为众数所在组的频数;f_{-1} 为众数所在组前一组的频数;f_{+1} 为众数所在组后一组的频数;d 为众数所在组的组距。

【例 5-4】 利用图 5-4 的资料计算 120 家企业利润额的众数。

在 Excel 中的操作步骤如下。

① 在单元格 C8 中输入"=MAX(C2:C7)",按 Enter 键,结果为 42,则该众数所在组为"400～500"这一组。

② 在单元格 E8 中输入"=400+(B4-B3)/((B4-B3)+(B4-B5))*100",按 Enter 键,即得众数为 433.33。

	A	B	C	D	E
1	按利润分组/万元		企业个数 f_i	向上累积频数	向下累积频数
2	300	以下	19	19	120
3	300	-400	30	49	101
4	400	-500	42	91	71
5	500	-600	18	109	29
6	600	以上	11	120	11
7	合计		120	-	-
8	众数所在组:		42	众数:	433.33

图 5-4　某地区 120 家企业利润额情况表

利用式(5-1)和式(5-2)计算众数时是假定数据分布具有明显的集中趋势,且众数组的频数在该组内是均匀分布的,若这些假定不成立,则众数的代表性就会很差。从众数的计算公式可以看出,众数是根据众数组及相邻组的频率分布信息来确定数据中心点位置的,因此,众数是一个位置代表值,它不受数据中极端值的影响。

二、中位数

中位数是一组数据按从小到大排序后,处于中间位置上的变量值,用 M_e 表示。显然,中位数将全部数据等分成两部分,有 50%的数据比中位数大,有 50%的数据则比中位

数小，所以，中位数可以反映出一组数据的中心位置、数据的集中趋势。当一组数据中含有极端值时，用中位数反映该组数据的一般水平可避免极端数值对平均数的影响。

在 Excel 中计算中位数的函数为 MEDIAN()，用于找出数据分布中心位置的数据。

其语法结构为：MEDIAN(number1，number2，...)。

其中，number1，number2，… 为需要找出中位数的 1～30 个参数，也可以是单一数组或者对数据区域的引用。如果一组数据中包含偶数个数字，函数 MEDIAN()将返回位于中间的两个数的平均值。

中位数主要用于顺序型数据的平均分析，也可以作为数值型数据的分析，但不适用于分类型数据。

(1) 对于顺序型数据，我们可以首先用公式$\frac{n+1}{2}$(n 为数据的个数)确定中位数的位置，然后根据向上累积频数确定中位数的具体类别。

【例 5-5】 根据第四章表 4-4 和表 4-5 中的数据，计算甲、乙两个城市对住房状况满意度的中位数。

在 Excel 中的操作步骤如下。

① 在某一空白单元格中输入公式"＝(300＋1)/2"，按 Ener 键，则结果为 150.5，即是两组数据中位数的位置所在。

② 从两表的向上累积频数可以看出，甲、乙两个城市的中位数在"一般"这一类别中，因此甲、乙两个城市的中位数 M_e 均为"一般"。

也就是说，对于甲、乙两个城市而言，可以用"一般"来代表家庭对住房状况的满意程度。当然，其代表性如何还需通过差异分析来解决。

(2) 对于未分组数据，我们可以直接利用 MEDIAN()函数确定中位数。

【例 5-6】 有甲、乙两组工人加工某种零件，日加工零件数如下(单位：个)，计算两组的中位数。

甲组：55　58　59　59　61　60　63　65

乙组：55　56　56　58　57　61　61　61　63

在 Excel 中的操作步骤如下。

① 在 Excel 中输入两组数据，如图 5-5 所示。

	A	B	C	D	E	F	G	H	I	J
1	甲组	55	58	59	56	61	60	63	65	
2	乙组	55	56	56	58	57	61	61	61	63

图 5-5　未分组数据的中位数

② 在某个空白单元格中输入公式："＝MEDIAN(A1:I1)"，按 Enter 键，结果为59.5，即是甲组的中位数。

③ 在另一个单元格中输入公式："＝MEDIAN(A2:J2)"，按 Enter 键，结果为 58，即是乙组的中位数。

由此可见，中位数是一个位置代表值，其特点是不受极端值的影响。中位数在研究收入分配时很有用。

(3) 对于单变量值分组数据,我们可以首先根据公式 $\frac{n}{2}$ 或 $\frac{\sum_{i=1}^{k} f_i}{2}$ 确定中位数所在组,然后找到该组的标志值即为中位数。

【例 5-7】 某车间工人按日产量的频数分布表如图 5-6 所示,确定其中位数。

	A	B	C	D
1	某车间工人的日产量频数分布表			
2	按日产量分组/个	工人数	向上累积频数	向下累积频数
3	55	6	6	40
4	56	8	14	34
5	58	13	27	26
6	59	10	37	13
7	61	3	40	3
8	合计	40	-	-

图 5-6 某车间工人的日产量频数分布表

在 Excel 中的操作步骤如下。

① 在某个空白单元格中输入公式"=40/2",按 Enter 键,结果为 20。

② 根据向上累积频数可以确定中位数所在组为 58 这一组,则该组的中位数为 58。

(4) 对于组距分组数据,我们可以首先根据公式 $\frac{n}{2}$ 或 $\frac{\sum_{i=1}^{k} f_i}{2}$ 确定中位数的位置,并确定中位数所在的组,然后采用下面的公式计算中位数的近似值。

下限公式:

$$M_e = L + \frac{\frac{\sum_{i=1}^{K} f_i}{2} - S_{m-1}}{f_m} \cdot d \tag{5-3}$$

上限公式:

$$M_e = U - \frac{\frac{\sum_{i=1}^{k} f_i}{2} - S_{m+1}}{f_m} \cdot d \tag{5-4}$$

式中,L 为中位数所在组的下限;U 为中位数所在组的上限;S_{m-1} 为中位数所在组前一组的累积频数;S_{m+1} 为中位数所在组后一组的累积频数;f_m 为中位数所在组的频数;$\sum_{i=1}^{K} f_i$ 为各组频数之和,即总频数;d 为中位数所在组的组距。

【例 5-8】 利用图 5-4 的资料计算 120 家企业利润额的中位数。

在 Excel 中的操作步骤如下。

① 在某个空白单元格中输入公式"=120/2",按 Enter 键,结果为 60。

② 根据向上累积频数可以确定中位数所在组为"400～500"这一组。

③ 在另一空白单元格中输入下限公式"=400+((C7/2)-D3))/C4*100",按 Enter 键,结果为 426.19。

利用分组资料计算中位数时，假定中位数所在组的频数在该组内是呈均匀分布的。由于中位数是一个位置代表值，其数值的大小不受极大值和极小值的影响，因此中位数具有稳健性或耐抗性的特点。此外，中位数还具有这样一个性质，就是各变量值与中位数的离差绝对值之和最小，即 $\sum_{i=1}^{n}|x_i - M_e| = \min$。这一性质表明，中位数与各数据的距离之和最短，这在工程设计中有一定的应用价值。

三、分位数

中位数是从中间点将全部数据等分为两部分。与中位数类似的还有四分位数、十分位数和百分位数等。它们分别是用 3 个点、9 个点和 99 个点将数据 4 等分、10 等分和 100 等分后各分位点上的值。这里只介绍四分位数的计算，其他分位数与之类似。

一组数据排序后处于 25％和 75％位置上的值，称为四分位数，也称四分位点。

四分位数是通过 3 个点将全部数据等分为四部分，其中每部分包含 25％的数据。很显然，中间的四分位数就是中位数，因此通常所说的四分位数是指处在 25％位置上的数值（下四分位数）和处在 75％位置上的数值（上四分位数）。与中位数的计算方法类似，根据未分组数据计算四分位数时，首先对数据进行排序，然后确定四分位数所在的位置。

设下四分位数为 Q_L，上四分位数为 Q_U，对于未分组的原始数据，各四分位数的位置分别为

$$Q_L \text{位置} = \frac{n+1}{4},\quad Q_U \text{位置} = \frac{3(n+1)}{4}$$

当四分位数的位置不在某一个位置上时，可根据四分位数的位置，按比例分摊四分位数两侧的差值。

在 Excel 中计算四分位数的函数为 QUARTILE()。

其语法结构为：QUARTILE(array，quart)。

其中，array 为计算四分位数的数组或数据区域；quart 决定返回第几个四分位数，quart＝0 时，QUARTILE 返回最小值；quart＝1 时，QUARTILE 返回第 1 四分位数；quart＝2 时，QUARTILE 返回中位数；quart＝3 时，QUARTILE 返回第 3 四分位数；quart＝4 时，QUARTILE 返回最大值。

【例 5-9】 根据例 5-6 中甲组工人的日产量数据，计算日产量的四分位数。

在 Excel 中的操作步骤如下。

（1）在某空白单元格中输入公式“＝QUARTILE($B$1:$I$1，0)”，按 Enter 键，结果为 55。

（2）用鼠标向下拖动四个单元格来复制公式。

（3）把复制的四个单元格中 QUARTILE()函数的第二个参数依次更改为 1，2，3，4，则结果依次为：57.5，59.5，61.5，65。

四、算术平均数

算术平均数也称为均值，是全部数据算术平均的结果。算术平均法是计算平均指标

最基本、最常用的方法。在社会经济现象中,有很多研究总体,其总体标志总量常常是总体单位变量值的算术总和。例如,工人工资总额是总体中每个工人工资的总和,某地区小麦总产量是所有耕地小麦产量的总和。在总体标志总量和总体单位总量的基础上,就可以计算平均指标。计算公式为

$$\text{算术平均数} = \frac{\text{总体标志总量}}{\text{总体单位总量}}$$

算术平均数在统计学中具有重要的地位,是平均分析的最主要度量值,通常用 $\bar{x}$(读作 x-bar)表示。它主要用于数值型数据,而不适用于分类数据和顺序型数据。根据所掌握数据形式的不同,算术平均数有简单算术平均数和加权算术平均数两种。

(一) 简单算术平均数

简单算术平均数适用于未分组整理的原始数据,其算术平均数的计算就是直接将一组数据的各个数值相加除以数值个数。

设一组数据为 $x_1, x_2, \cdots, x_n$,则算术平均数 $\bar{x}$ 的计算公式为

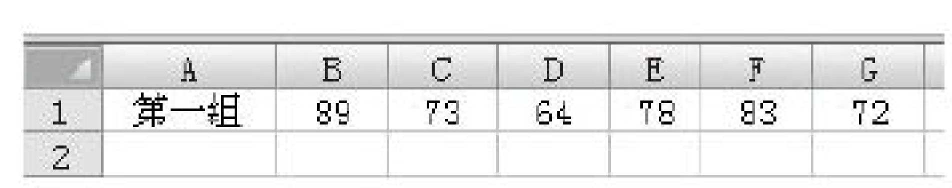
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5-5}$$

在 Excel 中计算算术平均数的函数为 AVERAGE(),它用于计算一组数值的算术平均数。

其语法结构为:AVERAGE(number1,number2,…)。

其中,number1,number2,…为需要计算算术平均数的 1~30 个参数,也可以是数组或单元格区域的引用。

【例 5-10】 某机械厂加工车间第一组有工人 6 名,某一日加工零件(单位:个)如图 5-7 所示,计算平均每人某一个工作日的产量。



	A	B	C	D	E	F	G
1	第一组	89	73	64	78	83	72
2							

图 5-7 某机械厂加工车间第一组日加工零件数

在 Excel 中的操作步骤如下。

在某空白单元格中输入公式"=AVERAGE(B1:G1)",按 Enter 键,结果为 76.5。即平均每人某一个工作日的产量约为 77 个。

(二)加权算术平均数

加权算术平均数适用于分组数据,以各组变量值出现的次数或频数为权数计算加权的算术平均数。

设原始数据被分成 k 组,各组的变量值为 $x_1, x_2, \cdots, x_k$,各组变量值的次数或频数分别为 $f_1, f_2, \cdots, f_k$,则加权的算术平均数为

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} \tag{5-6}$$

【例 5-11】 某汽车配件厂有 40 名工人,他们每人每日加工的某种零件数,编制成单

项数列资料如图 5-8 所示，计算 40 名工人平均每人每日加工零件数。

在 Excel 中的操作步骤如下：

(1) 在 C2 单元格中输入公式“＝A2 * B2”，按 Enter 键；把鼠标移动到 C2 单元格右下角，当变为“＋”形状时按住左键不放，向下拖曳至 C8 单元格，松开左键。

(2) 在 B9 单元格输入公式“＝SUM(B2:B8)”，按 Enter 键；把鼠标移动到 B9 单元格右下角，当变为“＋”形状时按住左键不放，向右拖曳至 C9 单元格，松开左键。

	A	B	C	D
1	按日产量分组 x_i/件	企业个数 f_i	$x_i * f_i$	算术平均数
2	25	1	25	29.075
3	26	2	52	
4	27	6	162	
5	29	12	348	
6	30	15	450	
7	31	2	62	
8	32	2	64	
9	合计	40	1 163	

图 5-8　某汽车配件厂工人每人每日加工某种零件情况

(3) 在 D2 单元格中输入公式“＝C9/B9”，按 Enter 键，即为企业利润额的加权算术平均数。

从例 5-11 可以看出，加权算术平均数其数值的大小，不仅受各组变量值(x_i)大小的影响，而且受各组变量值出现的频数(f_i)大小的影响。如果某一组的频数较大，说明该组的数据较多，那么该组数据的大小对算术平均数的影响就越大；反之，则越小。由于各组频数(f_i)对平均数的大小起到权衡轻重的作用，所以称各组频数为权数。实际上，我们将式(5-6)变形为下面的形式，就更能清楚地看出这一点。

$$\bar{x} = \frac{\sum_{i=1}^{K} x_i f_i}{\sum_{i=1}^{K} f_i} = \sum_{i=1}^{K} x_i \frac{f_i}{\sum_{i=1}^{K} f_i} \tag{5-7}$$

由式(5-7)可以清楚地看出，加权算术平均数受各组变量值(x_i)和各组权数即频率 $f_i \Big/ \sum_{i=1}^{K} f_i$ 大小的影响。当我们掌握的权数不是各组变量值出现的频数，而是频率时，可直接根据式(5-7)计算算术平均数。

例 5-11 是根据单变量值分组数据计算的加权算术平均数，如果根据组距分组数据计算平均数时，则要用各组的组中值代替变量值，然后进行加权平均。举例如下。

【例 5-12】 在某地区抽取 120 家企业，按利润额进行分组，如图 5-9 所示。现利用表中资料计算平均数。

	A B	C	D	E
1	按利润分组 /万元	组中值 x_i	企业个数 f_i	$x_i * f_i$
2	300 以下	250	19	4 750
3	300～400	350	30	10 500
4	400～500	450	42	18 900
5	500～600	550	18	9 900
6	600 以上	650	11	7 150
7	合计	2 250	120	51 200
8	加权平均数:	426.67		

图 5-9　某地区 120 家企业利润额情况表

在 Excel 中的操作步骤如下。

(1) 计算组中值。在 C2 单元格中输入公式“＝A2－100/2”，按 Enter 键；在 C3 单元格中输入“＝A3＋100/2”，按 Enter 键；把鼠标移动到 C3 单元格右下角，当变为“＋”形状时按住左键不放，向下拖曳至 C6 单元格，松开左键。

(2) 在 E2 单元格中输入公式“＝C2 * D2”，按 Enter 键；把鼠标移动到 E2 单元格右下角，当变为“＋”形状时按住左键不放，向下拖曳至 E6 单元格，松开左键。

(3) 在 C7 单元格输入公式"=SUM(C2:C6)",按 Enter 键;把鼠标移动到 C2 单元格右下角,当变为"**+**"形状时按住左键不放,向右拖曳至 E7 单元格,松开左键。

(4) 在 C8 单元格中输入公式"=E7/D7",按 Enter 键,即为企业利润额的加权算术平均数。

需要说明的是,用各组的组中值代替变量值计算具有假定性,即假定各组内部的变量值是均匀分布或对称分布的。事实上,这种均匀分布或对称分布的情况很少,因此,由组距数列计算的加权平均数只是平均数的近似值。

五、调和平均数

调和平均数又称倒数平均数,是总体各单位标志值倒数的算术平均数。习惯上用 H 表示。从计算方法来说,调和平均数分为简单调和平均数和加权调和平均数两种。

(一) 简单调和平均数

为了方便理解简单调和平均数的概念和计算方法,我们举一个简单的例子。

例如,某市场某种蔬菜的价格,早上 5.5 元/千克,中午 4.5 元/千克,下午 4 元/千克。若早、午、晚各买 1 千克蔬菜和各买 1 元钱的蔬菜,求蔬菜的平均价格。

对于第一种购买行为,实际上就是购买了 3 千克蔬菜,其平均价格的计算就用简单算术平均数公式,计算如下:

$$\text{蔬菜平均价格} = \frac{5.5+4.5+4}{1+1+1} = 4.67(\text{元 / 千克})$$

对于第二种购买行为,实际上就是购买了 3 元钱的蔬菜,要计算平均价格,首先应计算早午晚各买蔬菜的数量,然后才能计算出购买蔬菜的平均价格。

早上购买蔬菜的数量$=\frac{1}{5.5}=0.18$(千克);中午购买蔬菜的数量$=\frac{1}{4.5}=0.22$(千克);下午购买蔬菜的数量$=\frac{1}{4}=0.25$(千克)。

$$\text{蔬菜平均价格} = \frac{1+1+1}{\frac{1}{4.5}+\frac{1}{4}+\frac{1}{5.5}} = 4.62(\text{元 / 千克})$$

由于以上资料中缺乏总体单位总量,所以,不能直接用算术平均的方法计算平均指标。首先要用变量值的倒数计算出总体单位总量来,然后再计算平均指标。这种方法计算的平均数被称为简单调和平均数。计算公式为

$$H = \frac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}{n}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{5-8}$$

相对应的 Excel 调和平均数函数为 HARMEAN,它用于计算一组数值的调和平均数。

其语法结构为 HARMEAN(number1,number2,…)。

其中,number1,number2,…为需要计算调和平均数的 1～30 个参数。Excel 的操作

步骤与方法同算术平均数。

（二）加权调和平均数

在上例中，如果在早上花费5元，中午花费8元，下午花费10元，购买这些蔬菜的平均价格是多少？

$$\text{蔬菜平均价格}=\frac{5+8+10}{\frac{5}{5.5}+\frac{8}{4.5}+\frac{10}{4}}=4.43(\text{元}/\text{千克})$$

由上例可以看出，三个集市购买蔬菜所花费的现金在计算平均价格的过程中起到权数的作用，因此，我们把这种方法称为加权调和平均法。计算公式为

$$H=\frac{m_1+m_2+\cdots+m_k}{\frac{m_1}{x_1}+\frac{m_2}{x_2}+\cdots+\frac{m_k}{x_k}}=\frac{\sum_{i=1}^{k}m_i}{\sum_{i=1}^{k}\frac{m_i}{x_i}} \tag{5-9}$$

【例5-13】 某采购站某月购进四批同类商品，每批的单价及采购金额如图5-10所示，试计算四批产品的平均单价。

	A	B	C	D
1	采　购 批　次	价格/(元/千克) x_i	采购金额/元 m_i	采购量/千克 m_i/x_i
2	第一批	30	7 800	260
3	第二批	35	8 400	240
4	第三批	40	12 000	300
5	第四批	45	18 000	400
6	合　计	-	46 200	1 200

图5-10　某采购站某月购进商品情况

根据图5-10给出的原始数据（四批次商品的单价与采购额）计算平均价格时，无法直接采用加权算术平均数形式。这时，需要根据单价和采购额数据先求出采购量数据，再用总采购额除以总采购量即得平均价格，即加权调和平均。根据图5-10所示的数据，代入式(5-9)得平均价格为

$$H=\sum_{i=1}^{k}m_i\Big/\sum_{i=1}^{k}\frac{m_i}{x_i}=\frac{46\ 200}{1\ 200}=38.5(\text{元})$$

事实上，式(5-9)只是加权算术平均数的另一种表现形式，式中m_i（采购额）实际上是采购单价x_i与采购量f_i的乘积，即$m_i=x_if_i$，这从下面的式中可以清楚地看出来。

$$H=\frac{\sum_{i=1}^{k}m_i}{\sum_{i=1}^{k}\frac{m_i}{x_i}}=\frac{\sum_{i=1}^{k}x_if_i}{\sum_{i=1}^{k}\frac{x_if_i}{x_i}}=\frac{\sum_{i=1}^{k}x_if_i}{\sum_{i=1}^{k}f_i}=\bar{x}$$

由此可见，调和平均数和算术平均数在本质上是一致的，唯一的区别是计算时使用了不同的数据。在实际应用时，可掌握这样的原则，当分子资料未知时，就采用加权算术平均数计算平均数，分母资料未知时，就采用加权调和平均数计算平均数。

六、几何平均数

几何平均数是n个变量值乘积的n次方根。在社会经济现象中，总体标志总量有时不等于总体各标志值的总和，而等于各标志值的乘积，这种情况下，只有几何平均数才能反映数量标志的平均水平。在实际应用中，通常用来计算平均比率或平均速度。

根据所掌握资料的表现形式不同,几何平均数可分为简单几何平均数和加权几何平均数两种形式。

(一) 简单几何平均数

简单几何平均数是 n 个变量乘积的 n 次方根,其计算公式为

$$G = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{5-10}$$

式中,G 为几何平均数,Π 为连乘符号。

在 Excel 中计算几何平均数的函数为 GEOMEAN(),它用于计算一组数值的几何平均数。

其语法结构为:GEOMEAN(number1,number2,…)。

其中,number1,number2,…为需要计算几何平均数的 1~30 个数值,也可以是数组或单元格区域的引用。

Excel 的操作步骤与方法同算术平均数。

【例 5-14】 某企业生产某种产品须经过毛坯、粗加工、精加工、装配四个连续作业车间才能完成,若某月份每个车间的产品合格率分别为 90%、96%、93%、87%,求该产品的企业合格率(四个车间的平均合格率)是多少。

由于产品是由四个车间连续加工完成的,每一个车间加工的都是前一个车间完工的合格制品,因此,四个车间总合格率是四个车间相应合格率的连乘积,求平均合格率就不能采用算术平均法,而应当用几何平均法。则四个车间平均合格率为

$$G = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} x_i^{f_i}} = \sqrt[4]{90\% \times 96\% \times 93\% \times 87\%} = 91.44\%$$

(二) 加权几何平均数

加权几何平均数的计算公式为

$$G = \sqrt[\sum_{i=1}^{k} f_i]{x_1^{f_1} x_2^{f_2} \cdots x_k^{f_k}} = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} x_i^{f_i}} \tag{5-11}$$

【例 5-15】 某地区 GDP 2005—2010 年平均发展速度为 107.2%,2011—2013 年平均发展速度为 108.7%,2014—2015 年平均发展速度为 110%,求该地区 2005—2015 年间的平均发展速度。

由于总速度是各年发展速度连乘形成的,该资料提供的各时段的平均发展速度所代表的时间长度又有所不同,所以根据该资料求平均发展速度需用加权几何平均法。2005—2015 年平均发展速度为

$$G = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} x_i^{f_i}} = \sqrt[10]{1.072^5 \times 1.087^3 \times 1.1^2} = 1.082(\text{或 } 108.2\%)$$

在变量值比较多的情况下,计算几何平均数需要开多次方。为简化计算,通常应用对数来进行计算。分别对式(5-10)和式(5-11)两端取对数,可得到以下两公式

$$\lg G=\frac{1}{n}(\lg x_1+\lg x_2+\cdots+\lg x_n)=\frac{\sum_{i=1}^{n}\lg x_i}{n}$$

$$\lg G=\frac{f_1\lg x_1+f_2\lg x_2+\cdots+f_n\lg x_n}{f_1+f_2+\cdots+f_n}=\frac{\sum_{i=1}^{n}f_i\lg x_i}{\sum_{i=1}^{n}f_i}$$

由此可见,几何平均数的对数等于各变量值对数的算术平均数,所以,几何平均数也称为对数平均数。

七、平均分析指标之间的关系

(一)算术平均数、调和平均数和几何平均数的关系

算术平均数、调和平均数和几何平均数都是数值平均数,即都是根据所有数据计算的。如果从纯数量关系上考察,这三种平均数的关系如下。

(1) 当一组数据中所有数据不尽相同时,据此计算的三种平均数的结果为:算术平均数最大,调和平均数最小,几何平均数居中。它们的关系用公式表示即为:$\overline{X}>G>H$。当一组数据中出现极端值时,通过这种关系我们不难看出,极端值对这三种平均数的影响程度是有差别的,它对算术平均数的影响最大,对几何平均数的影响次之,对调和平均数的影响最小。

(2) 当一组数据中所有的数据都相同时,据此计算的三种平均数相等,即$\overline{X}=G=H$。

三种数值平均数的这种关系是纯数学意义上的。当然,在实际应用中,采用何种平均数应取决于现象的客观性质和研究目的。就是说,适宜用算术平均数计算的,就不能用调和平均数或几何平均数计算,反之亦然。算术平均数是应用最为广泛的一种平均数,因为其计算方法是与许多社会经济现象的数量关系相符合的,即许多社会经济现象总体各单位的标志值之和等于总体的标志总量,且这种方法易理解并具有优良的数学性质。调和平均数在实际应用中,通常是作为算术平均数的变形使用的,即利用调和平均数的形式来计算算术平均数。几何平均数适合于对一些特殊数据,如比率、速度等的平均。

(二)算术平均数与众数、中位数的关系

算术平均数与众数、中位数的关系取决于频数分布的状况。它们的关系如下。

(1) 当数据具有单一众数且频数分布对称时,算术平均数与众数、中位数三者完全相等,即 $\overline{X}=M_o=M_e$,这一关系如图 5-11 所示。

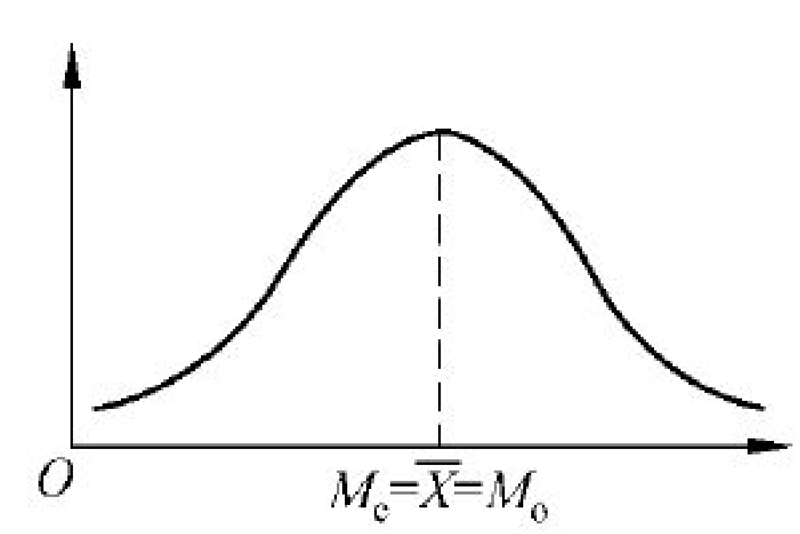


图 5-11 平均分析指标间关系

(2) 当频数分布呈现右偏态时,说明数据存在最大值,必然拉动算术平均数向极大值一方靠,则三者之间的关系为 $\overline{X}>M_e>M_o$,这一关系如图 5-12 所示。

(3) 当频数分布呈现左偏态时,说明数据存在最小

值,必然拉动算术平均数向极小值一方靠,而众数和中位数由于是位置平均数,不受极值的影响,因此,三者之间的关系为 $\overline{X}<M_e<M_o$,这一关系如图 5-13 所示。

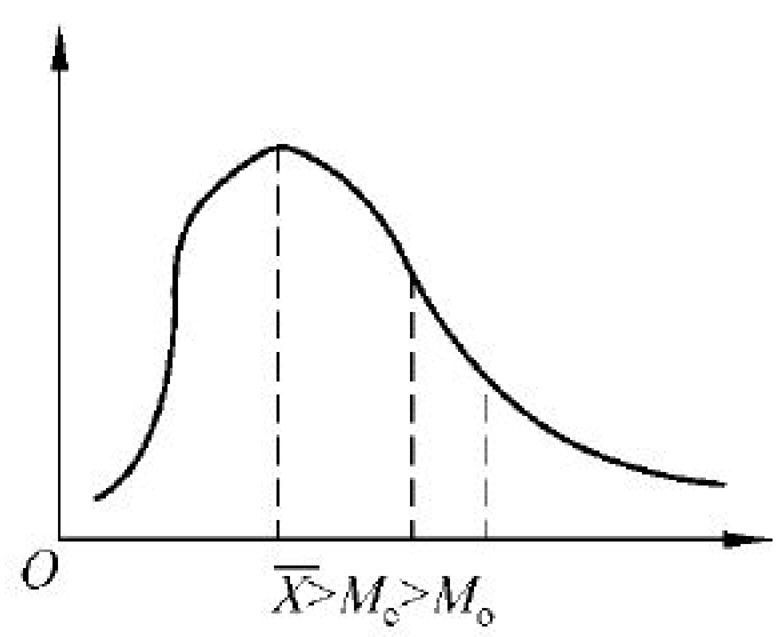


图 5-12 平均分析指标间关系

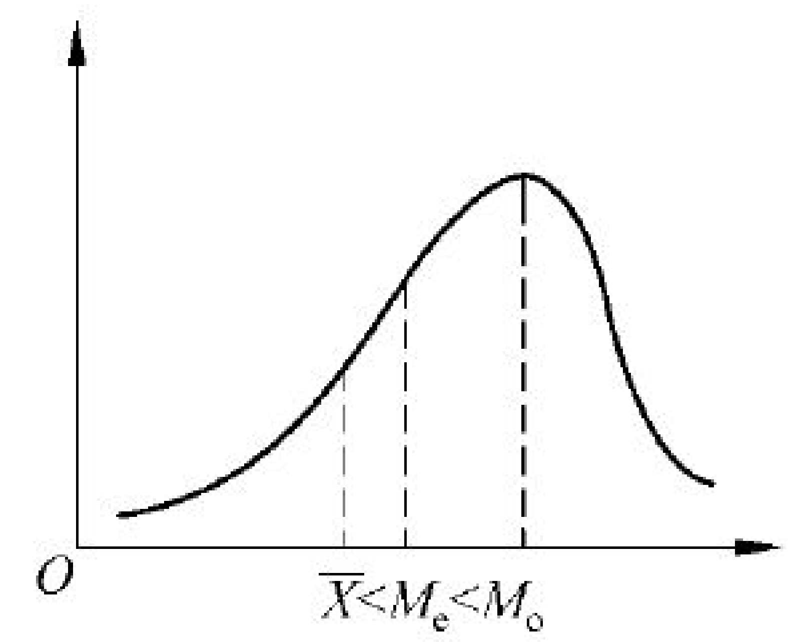


图 5-13 平均分析指标间关系

从上面的分析我们可以看出,当频数分布出现偏态时,极端值对算术平均数产生很大的影响,而对众数、中位数没有影响,此时,用众数、中位数作为一组数据的中心值比算术平均数有较高的代表性。

第三节 差异分析

根据所依据数据类型的不同,差异分析指标有异众比率、四分位差、全距、平均差、方差和标准差、离散系数等。

一、异众比率

非众数组的频数占总频数的比率,称为异众比率,用 V_r 表示。

计算公式为

$$V_r=\frac{\sum f_i-f_m}{\sum f_i}=1-\frac{f_m}{\sum f_i} \tag{5-12}$$

式中,$\sum f_i$ 为变量值的总频数;f_m 为众数组的频数。

异众比率的作用是衡量众数对一组数据的代表性程度的指标。异众比率越大,说明非众数组的频数占总频数的比重就越大,众数的代表性就越差;反之,异众比率越小,众数的代表性就越好。

【例 5-16】 一家调查公司为了了解不同手机品牌的市场占有率,对随机抽取的一家手机零售商场进行调查,调查员在某天对 100 名顾客购买手机的品牌进行了记录,记录数据经分类整理后形成频数分布表,如图 5-14 所示,试计算异众比率。

	A	B	C
1	手机品牌	频数	频率
2	诺基亚	40	40%
3	摩托罗拉	23	23%
4	三星	20	20%
5	联想	10	10%
6	其他	7	7%
7	合计	100	1

图 5-14 调查数据频数分布情况

在 Excel 中的操作步骤如下。

在某空白单元格中输入公式:"=1-(40/100)",按 Enter 键,结果为 60%。

该资料表明,在所调查的 100 人当中,购买诺基亚手机的人数最多,众数为 40,购买诺基亚之外的其他手机品牌的人数占 60%,异众比率比较大。因此,用"诺基亚"来反映

居民对手机品牌偏好的一般趋势，其代表性不是很好。

二、四分位差

四分位差，也称为内距或四分间距，它是上四分位数与下四分位数之差，用 Q_d 表示。计算公式为

$$Q_d = Q_U - Q_L \tag{5-13}$$

四分位差反映了中间 50%数据的差异程度，其数值越小，说明中间的数据越集中；数值越大，说明中间的数据越分散。此外，由于中位数处于数据的中间位置，因此，四分位差的大小在一定程度上也说明了中位数对一组数据的代表程度。四分位差不受极端值的影响。

【例 5-17】 根据例 5-9 的资料计算上下四分位数，计算工人日产量的四分位差为

$$Q_d = Q_U - Q_L = 61.5 - 57.5 = 4(\text{个})$$

三、全距

全距又称极差，是一组数据的最大值与最小值之差，用 R 表示。计算公式为

$$R = \max(X_i) - \min(X_i) \tag{5-14}$$

式中，$\max(X_i)$，$\min(X_i)$分别为一组数据的最大值与最小值。

由于全距是根据一组数据的两个极值表示的，所以全距表明了一组数据数值的变动范围。R 越大，表明数值变动的范围越大，即数列中各变量值差异大；反之，表明数值变动的范围越小，即数列中各变量值差异小。

【例 5-18】 根据例 5-9 中的数据，可计算甲组工人日产量的全距为

$$R = 65 - 55 = 10(\text{个})$$

全距是描述差异程度的最简单度量值，计算简单直观，易于理解，但其数值大小易受极端值的影响。由于全距只是利用了一组数据两端的信息，不能反映中间变量值的差异，因而不能准确描述出数据的差异程度。

四、平均差

平均差也称平均绝对离差，它是各变量值与其算术平均数离差绝对值的平均数，用 M_D 表示。根据掌握资料的不同，平均差有以下两种计算方法。

(1) 对于未分组资料，采用简单平均法。其计算公式为

$$M_D = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n} \tag{5-15}$$

在 Excel 中可以利用 AVEDEV()函数计算一组数值型数据的平均差。

其语法结构为：AVEDEV(number1,number2,…)。

其中，number1，number2，…为需要计算平均差的 1～30 个数值，也可以是数组或单元格区域的引用。

【例 5-19】 根据图 5-4 中的数据，计算两组工人日产量的平均差。

在 Excel 中的操作步骤如下。

① 在某空白单元格中输入公式"=AVEDEV(B1:I1)",按 Enter 键,结果为 2.63,即甲组工人日产量的平均差。

② 在某空白单元格中输入公式"=AVEDEV(B2:J2)",按 Enter 键,结果为 2.52,即乙组工人日产量的平均差。

从计算结果可以看出,甲组工人日产量的差异比乙组工人的差异大。

(2) 对于分组资料,应采用加权平均法,其计算公式为

$$M_D = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| f_i}{\sum_{i=1}^{k} f_i} \tag{5-16}$$

【例 5-20】 利用例 5-12 中的资料,计算 120 家企业利润额的平均差。

在 Excel 中的计算过程和结果如图 5-15 所示。(具体操作步骤略)

	A B	C	D	E	F
1	按利润分组/万元	组中值 x_i	企业个数 f_i	$\|x_i - \bar{x}\|$	$\|x_i - \bar{x}\| * f_i$
2	300以下	250	19	176.67	3 356.67
3	300~400	350	30	76.67	2 300.00
4	400~500	450	42	23.33	980.00
5	500~600	550	18	123.33	2 220.00
6	600以上	650	11	223.33	2 456.67
7	合计	2 250	120	623.33	11 313.33
8	均值	426.666 7	平均差		94.28

图 5-15 分组数据平均差的计算

计算结果表明,该地区 120 家企业利润额的差异程度平均为 94.28 万元。

平均差计算简便,意义明确,而且平均差是根据所有变量值计算的,因此它能够准确、全面地反映一组数值的变异程度。平均差越大,说明数据的离散程度越大;反之,说明数据的离散程度越小。但是,由于平均差是用绝对值进行运算的,这就给计算带来了不便,所以在实际工作中运用较少。

五、方差和标准差

方差是各变量值与其算术平均数离差平方的算术平均数。标准差是方差的平方根。

方差和标准差同平均差一样,也是根据全部数据计算的,反映每个数据与其算术平均数相比平均相差的数值,因此它能准确地反映出数据的差异程度。与平均差相比,方差与标准差在数学处理上是通过平方消除离差的正负号,更便于数学上的处理。因此,方差、标准差是实际中应用最广泛的离中程度度量值。根据总体数据和样本数据计算的方差和标准差在数学处理上略有不同。

(一) 总体的方差和标准差

总体的方差一般用符号 σ^2 表示,标准差一般用符号 σ 表示。

(1) 对于未分组整理的原始资料,方差和标准差的计算公式分别为

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} \tag{5-17}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}} \tag{5-18}$$

在 Excel 中可以利用 VARP()函数计算一组数值型数据的总体方差;利用 STDEVP()函数计算一组数值型数据的总体标准差。

其语法结构分别为:VARP(number1,number2,…),

STDEVP(number1,number2,…)。

其中,number1,number2,…为需要计算平均差的 1～30 个数值,也可以是数组或单元格区域的引用。

【例 5-21】 假设图 5-4 中的数据是甲、乙两个总体的全部日产量数据,计算甲组工人日产量的方差和标准差。

在 Excel 中的操作步骤如下。

① 在某空白单元格中输入公式"=VARP(B1:I1)",按 Enter 键,结果为 9.98,即甲组工人日产量的方差。

② 在某空白单元格中输入公式"=STDEVP(B1:I1)",按 Enter 键,结果为 3.16,即甲组工人日产量的标准差。

(2) 对于分组数据,方差和标准差的计算公式分别为

$$\sigma^2 = \frac{\sum_{i=1}^{K}(X_i - \overline{X})^2 F_i}{\sum_{i=1}^{K} F_i} \tag{5-19}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K}(X_i - \overline{X})^2 F_i}{\sum_{i=1}^{K} F_i}} \tag{5-20}$$

【例 5-22】 假设例 5-12 中的资料为总体资料,计算 120 家企业利润额的方差和标准差。

在 Excel 中计算过程和结果如图 5-16 所示。(具体操作步骤略)

	A	B	C	D	E	F
1	按利润分组/万元		组中值 x_i	企业个数 f_i	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2*f_i$
2	300以下		250	19	31 211.11	593 011.11
3	300～400		350	30	5 877.78	176 333.33
4	400～500		450	42	544.44	22 866.67
5	500～600		550	18	15 211.11	273 800.00
6	600以上		650	11	49 877.78	548 655.56
7	合计		2 250	120	102 722.22	1 614 666.67
8				方差		13 455.56
9	均值		426.666 7	标准差		116.00

图 5-16　分组数据方差、标准差的计算

与方差不同的是,标准差是具有量纲的,它与变量值的计量单位相同,其实际意义要比方差清楚。因此,在对实际问题进行分析时,我们更多地使用标准差。

(二)样本的方差和标准差

样本的方差和标准差与总体的方差和标准差在计算上有所差别。总体的方差和标准差在对各个离差平方平均时是除以数据个数或总频数,而样本的方差和标准差在对各个离差平方平均时是用样本数据个数或总频数减 1 去除总离差平方和。

样本的方差为一般用符号“s^2”表示,样本的标准差一般用 s 表示。

(1) 对于未分组整理的原始资料,方差和标准差的计算公式为

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{5-21}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{5-22}$$

在 Excel 中可以利用 VAR()函数计算一组数值型数据的样本方差;利用 STDEV()函数计算一组数值型数据的样本标准差。

其语法结构分别为:VAR(number1,number2,…),

STDEV(number1,number2,…)。

其中,number1,number2,…为需要计算平均差的 1~30 个数值,也可以是数组或单元格区域的引用。

【例 5-23】 假设图 5-5 中的数据是日产量的样本数据,计算甲组工人日产量的方差和标准差。

在 Excel 中的操作步骤如下。

①在某空白单元格中输入公式“=VAR(B1:I1)”,按 Enter 键,结果为 11.4,即甲组工人日产量的方差。

②在某空白单元格中输入公式“=STDEV(B1:I1)”,按 Enter 键,结果为 3.38,即甲组工人日产量的标准差。

(2) 对于分组数据,方差和标准差的计算公式分别为

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1} \tag{5-23}$$

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1}} \tag{5-24}$$

【例 5-24】 假设例 5-12 中的资料为样本资料,120 家企业利润额的方差和标准差为

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1} = \frac{1\ 614\ 666.67}{120 - 1} = 13\ 568.63$$

$$s = \sqrt{13\ 568.63} = 116.48(\text{万元})$$

这与根据总体的方差和标准差计算公式计算的结果相差不大。当 n 很大时，样本方差 s^2 与总体方差 σ^2 的计算结果相差很小，这时样本方差也可以用总体方差的公式来计算。

（三）是非标志的方差和标准差

在社会经济统计中，有些事物或现象的特征只表现为两种性质上的差异。例如，产品的质量表现为合格或不合格，人的性别表现为男或女，对某一电视节目表现为收看或不收看，等等。这些只表现为是与否、有或无的标志，称为是非标志，也称为交替标志。在进行抽样估计时，是非标志的方差或标准差具有很重要的意义。

1. 成数（比例）

成数是非标志只有两种表现，我们把总体中或样本中具有某种表现或不具有某种表现的单位数占全部单位数的比重称为成数，它反映了总体或样本中“是”与“非”的构成，并且代表着两种表现或性质各反复出现的程度，即频率。例如，某一批产品，合格品占95%，不合格品占5%。在这里，95%和5%均为成数。

若以 N_1 表示总体中具有某种表现的单位数，N_0 表示总体中不具有某种表现的单位数，N 表示总体单位数，则成数可表示为

$$\pi = \frac{N_1}{N} \quad \text{或} \quad 1 - \pi = \frac{N_0}{N}$$

对于样本来说，与总体 N_1 对应的就是 n_1，与总体 N_0 对应的就是 n_0，样本单位数为 n，则有

$$p = \frac{n_1}{n} \quad \text{或} \quad 1 - p = \frac{n_0}{n}$$

在同一总体或样本中，对于某一交替标志，具有两种成数且其和为1。

2. 是非标志的平均数

是非标志是一种品质标志，其表现为文字。因此，在计算平均数时，首先需要将文字表现进行数量化处理。用“1”表示具有某种表现，用“0”表示不具有某种表现，然后以“1”和“0”作为变量值，计算加权算术平均数。现以总体为例予以说明。

$$\overline{X}_n = \frac{\sum X_i F_i}{\sum F_i} = \frac{1 \times N_1 + 0 \times N_0}{N_1 + N_0} = \frac{N_1}{N} = \pi \tag{5-25}$$

可见，总体是非标志的平均数，即为被研究标志具有某种表现的成数 π，同样可得样本是非标志的平均数即为被研究标志具有某种表现的成数 p。

3. 是非标志的方差与标准差

将经过量化处理的是非标志的表现“1”和“0”作为变量值代入总体的方差计算公式

$$\sigma_\pi^2 = \frac{\sum (X_i - \overline{X})^2 F_i}{\sum F_i} = \frac{(1-\pi)N_1 + (0-\pi)^2 N_0}{N_1 + N_0} = \pi(1-\pi)$$

为区别于一般变量值的方差,我们将是非标志的方差记为 σ_π^2,即

$$\sigma_\pi^2 = \pi(1-\pi) \tag{5-26}$$

是非标志的标准差为

$$\sigma_\pi = \sqrt{\pi(1-\pi)} \tag{5-27}$$

类似地,可得样本是非标志的方差 s^2 和标准差 s 为

$$s_p^2 = p(1-p) \tag{5-28}$$

$$s_p = \sqrt{p(1-p)} \tag{5-29}$$

【例 5-25】 某鞋厂生产皮鞋 5 000 双,从中抽取 100 双进行质量检查,结果为 98 双合格,2 双不合格,试计算合格品率的方差和标准差。

根据所给资料可得

$$p = \frac{98}{100} \times 100\% = 98\%, \quad 1-p = \frac{2}{100} \times 100\% = 2\%$$

$$s_p^2 = 98\% \times 2\% = 1.96\%, \quad s_p = \sqrt{1.96\%} = 14\%$$

当成数 $p=0.5$ 时,方差、标准差均为最大值,分别为 0.25 和 0.5,也就是说,此时是非标志的差异程度最大。如某学生群体中男生数和女生数相等,即成数 $p=0.5$(或 50%),说明该学生群体性别差异程度最大。

(四)相对位置的度量:标准分数

平均数是最为广泛使用的数据位置的度量,而方差和标准差是最为广泛使用的衡量数据差异程度的度量。利用均值和标准差,我们还可以计算一组数据中各个数值的标准分数,以测度每个数据在该组数据中的相对位置,并可以用它来判断一组数据是否有离群值。

变量值与其平均数的离差除以标准差后的值,称为标准分数,也称标准化值或 z 分数。计算公式为

$$z_i = \frac{x_i - \mu}{\sigma} \quad 或 \quad z_i = \frac{x_i - \overline{x}}{s} \tag{5-30}$$

z_i 可以解释为 x_i 距离平均数 $\overline{X}$ 的标准差的个数。比如,$z_i=-2$,我们就知道 x_i 比平均数小两个标准差。由式(5-30)得知,z 分数大于零表示相应的数据项比平均数大,z 分数小于零表示相应的数据项比平均数小,z 分数等于零表示相应的数据项等于平均数。

式(5-30)也就是我们常用的统计标准化公式,在对多个具有不同量纲的变量进行处理时,常常需要对各变量数值进行标准化处理。

在 Excel 中可以用 STANDARDIZE()函数计算某数值型数据的标准分数。

其语法结构为:STANDARDIZE(x,mean,standard_dev)

其中,x 为需要进行标准化的数值;mean 为该组数据的算术平均值;standard_dev 为

该组数据的标准差。

【例 5-26】 根据例 5-6 的数据，计算乙组每个工人日产量的标准分数。

在 Excel 中的操作步骤如下。（最终结果如图 5-17 所示）

	A	B	C	D	E	F	G	H	I	J
1	工人编号	1	2	3	4	5	6	7	8	9
2	日产量	55	56	56	58	57	61	61	61	63
3	标准分数	-1.277	-0.928	-0.928	-0.232	-0.580	0.812	0.812	0.812	1.509
4		日产量的均值:		58.667		日产量的标准差:		2.872		

图 5-17　工人日产量的标准分数

（1）计算日产量的均值：在 D4 单元格中输入公式："＝average(B2:J2)"，按 Enter 键，结果为 58.667。

（2）计算日产量的标准差：在 H4 单元格中输入公式："＝stdev(B2:J2)"，按 Enter 键，结果为 2.872。

（3）计算标准分数：在 B3 单元格中输入公式："＝(B2－\$D\$4)/\$H\$4"，按 Enter 键；再把鼠标移到 B3 单元格的右下角，按住鼠标左键不动，向右拖动至 J3 单元格，松开鼠标，即可得到各工人日产量的标准分数。

实际上，z 分数只是将原始数据进行了线性变换，它并没有改变一个数据在该组数据中的位置，也没有改变该组数分布的形状，而只是将该组数据变为均值为 0、标准差为 1。

经验法则表明，当一组数据对称分布时：

（1）约有 68%的数据在平均数加减 1 个标准差的范围内。

（2）约有 95%的数据在平均数加减 2 个标准差的范围内。

（3）约有 99%的数据在平均数加减 3 个标准差的范围内。

由此可见，一组数据中低于或高于平均数 3 个标准差以上的数据很少。几乎所有的数据项与平均数的距离都在 3 个标准差之内。因此，在统计上，往往将平均数 3 个标准差以外的数据称为异常值或离群值(outlier)。

六、离散系数

方差和标准差都是反映一组数值差异程度的绝对值，其数值的大小，不仅取决于数值的差异程度，而且还与变量值水平的高低、计量单位的不同有关。所以，不宜直接利用上述变异指标对不同水平、不同计量单位的现象进行比较，应当先做无量纲化处理，即将上述反映数据的绝对差异程度的变异指标转化为反映相对差异程度的指标，然后再进行对比。

离散系数也称为变异系数，它是一组数据的标准差与其相应的平均数的比值，是测度数据差异程度的相对指标。离散系数通常是用标准差来计算的，因此也称为标准差系数，其计算公式为

$$V_\sigma = \frac{\sigma}{\bar{X}} \quad 或 \quad V_s = \frac{s}{\bar{x}} \tag{5-31}$$

离散系数主要是用于比较不同总体或样本数据的差异程度。离散系数大的说明该组数据的差异程度也大，离散系数小的说明该组数据的差异程度也小。

【例 5-27】 已知甲、乙两个企业工人劳动生产率的资料如表 5-1 所示,计算并比较其离散系数。

表 5-1 甲乙两企业的工人劳动生产率

企业	工人平均劳动生产率/(元/人)	标准差/元	离散系数
甲	64 000	2 400	3.75%
乙	32 000	1 600	5.00%

计算结果表明,虽然从标准差来看,乙企业劳动生产率的差异比甲企业的小,但其离散系数却较大,所以乙企业劳动生产率的差异程度大于甲企业劳动生产率的差异程度。

本章介绍了平均分析的各种分析指标,其中大多数都可以通过 Excel 中的"数据分析"工具中的"描述统计"命令一次性得出计算结果。"数据分析"功能在 Excel 2013 中的调用及具体使用方法已在第四章中做了详细介绍,在此不再赘述。

思 考 题

1. 试述平均分析的含义及作用。
2. 平均分析的种类有哪些?
3. 简述众数、中位数和平均数的特点和应用场合。
4. 简述异众比率、四分位差、方差和标准差的计算方法及适用场合。
5. 标准分数的作用有哪些?
6. 为什么要计算离散系数?

练 习 题

1. 已知某企业工人的日产量及相应的人数资料如表 5-2 所示。

表 5-2 工人日产量及相应人数资料

日产量/(件/人)	72	68	65	70	69
工人数/人	9	23	20	16	11

要求:在 Excel 中计算该企业工人的平均日产量。

2. 某地区 2015 年的小麦产量分组资料如表 5-3 所示。

表 5-3 某地区 2015 年小麦产量分组资料

亩产量/(千克/亩[①])	播种面积比重/%	亩产量/(千克/亩)	播种面积比重/%
250 以下	10	3 000～3 500	40
2 500～3 000	30	3 500 以上	20

① 1 亩=$666.\dot{6}$ 平方米。

要求：在 Excel 中计算该地区小麦的平均产量。

3. 将一笔钱存入银行，存期为 10 年，按复利计息。10 年的利率分别是：第一年和第二年为 0.05，第三年至第五年为 0.08，第六年至第八年为 0.1，第九年和第十年为 0.12。求平均年利率。

4. 某水泥企业 2010 年的水泥产量为 100 万吨，2012 年与 2010 年相比增长率为 18%，2013 年与 2002 年相比增长率为 15%，2015 年与 2013 年相比增长率为 28%，求各年的平均增长率。

5. 某企业生产某产品的资料如表 5-4 所示。

表 5-4　某企业生产某产品资料表

日产量/件	工人数/人		日产量/件	工人数/人	
	第一天	第二天		第一天	第二天
400 以下	5	5	600 以上	10	9
400～500	15	6	合计	50	50
500～600	20	30			

要求：

(1) 在 Excel 中分别计算两天的平均每人生产的产品数量。

(2) 解释第二天比第一天生产率高的原因。

6. 某地区 120 家企业月产值的分组资料如表 5-5 所示。

表 5-5　120 家企业月产值的分组资料

月产值/万元	企业数/个	月产值/万元	企业数/个
100 以上	10	160～180	12
100～120	20	180 以上	10
120～140	48	合计	120
140～160	20		

要求：

(1) 在 Excel 中计算众数、中位数。

(2) 在 Excel 中计算平均差、方差、标准差。

7. 已知某百货公司 9 月各天的销售数据如表 5-6 所示。

表 5-6　某百货公司 9 月各天的销售数据　　万元

265	254	269	258	220	236	265	262	245	283
272	236	210	220	236	245	236	269	259	246
301	310	324	298	269	279	296	287	279	298

要求：(在 Excel 中计算如下指标)

(1) 计算该百货公司日销售额的平均数和中位数。

(2) 计算四分位数。

(3) 计算日销售额的标准差。

8. 有甲、乙两个生产小组,每组都是 8 个工人,各人的日产量件数如表 5-7 所示。

表 5-7 甲、乙生产小组各人的日产量件数

甲组	40	60	70	80	90	100	110	120
乙组	60	68	70	78	80	82	88	90

要求:

(1) 分别计算两组工人日产量的全距、平均差、标准差。

(2) 比较甲、乙两组的日平均产量代表性大小。

9. 两种不同的水稻分别在五块试验田上种植,其产量资料如表 5-8 所示。

表 5-8 产量资料表

试验田编号	甲品种		乙品种	
	试验田面积/亩	产量/斤①	试验田面积/亩	亩产量/斤
1	1.2	1 200	1.5	1 120
2	1.1	1 045	1.3	1 000
3	1.0	1 100	1.3	900
4	0.9	810	1.0	1 208
5	0.8	840	0.9	700
合计	5.0	4 995	6.0	—

要求:

(1) 计算两个品种的平均亩产量。

(2) 假定生产条件相同,分析哪一个品种具有较大的稳定性和推广价值?

10. 某企业有两甲、乙个生产车间生产三种产品的单位成本和总成本如表 5-9 所示。

表 5-9 单位成本和总成本资料

品种	单位成本/(元/件)	总成本/元		品种	单位成本/(元/件)	总成本/元	
		甲车间	乙车间			甲车间	乙车间
A	15	2 100	3 225	C	30	1 500	1 500
B	20	3 000	1 500				

要求:(在 Excel 中操作)

(1) 计算两个车间总平均成本的高低。

(2) 分析两个车间总平均成本差异的原因。

11. 某公司员工的工资情况如表 5-10 所示。

① 1 斤=0.5 千克。

表 5-10　某公司员工的工资情况表

月工资/(元/人)	员工人数/人	月工资/(元/人)	员工人数/人
800～1 000	5	1 400～1 600	10
1 000～1 200	10	1 600～1 800	5
1 200～1 400	20		

(1) 在众数、中位数和均值三个统计量中,你认为哪个统计量能概括该公司员工工资的情况? 为什么?

(2) 计算你认为能概括该公司员工工资情况的统计量。

12. 一家公司在招收职员时,首先要通过两项能力测试。在 A 项测试中,平均分数是 100 分,标准差是 15 分;在 B 项测试中,平均分数是 400 分,标准差是 50 分。一位应试者在 A 项测试中得了 115 分,在 B 项测试中得了 425 分。与平均分数相比,该位应试者哪一项测试更为理想?

13. 甲、乙两班各有 10 名同学参加一场智力测验,其成绩如表 5-11 所示。

表 5-11　甲、乙两班智力测验成绩表　分

甲班	98	89	87	78	95	76	93	92	93	62
乙班	91	87	94	96	91	93	97	88	91	86

用 Excel 对表 5-11 中的数据进行描述统计的结果如表 5-12 所示。

表 5-12　描述统计结果

统计量	甲　班	乙　班	统计量	甲　班	乙　班
平均成绩/分	86.3	91.4	极差	36.0	11.0
标准误差	3.5	1.2	最高成绩/分	62.0	86.0
中位数/分	90.5	91.0	最低成绩/分	98.0	97.0
众数/分	93.0	91.0	总成绩/分	863.0	914.0
标准差	11.1	3.7	人数/人	10.0	10.0
样本方差	123.1	13.6			

根据以上资料,你认为哪个班的同学考得好些? 为什么?

14. 为研究青少年儿童的成长发育状况,某研究所的一位调查人员在某城市抽取 100 名 7～17 岁的少年儿童作为样本,另一位调查人员则抽取了 1 000 名 7～17 岁的少年儿童作为样本。请回答下面的问题并解释其原因。

(1) 两位调查人员所得到的样本的平均身高是否相同? 如果不同,哪组样本的平均身高较大?

(2) 两位调查人员所得到的标准差是否相同? 如果不同,哪组样本的标准差较大?

(3) 两位调查人员得到这 1 100 名少年儿童身高的最高者或最低者的机会是否相同? 如果不同,哪位调查人员的机会较大?

第六章

时间序列分析

【统计知识 ABC】

［A］CPI

CPI(consumer price index，居民消费价格指数)指在反映一定时期内居民所消费商品及服务项目的价格水平变动趋势和变动程度。居民消费价格水平的变动率在一定程度上反映了通货膨胀(或紧缩)的程度。通俗地讲，CPI就是市场上的货物价格增长百分比。一般市场经济国家认为CPI增长率在2%～3%属于可接受范围内，当然还要看其他数据。CPI过高始终不是好事，高速经济增长率会拉高CPI，但物价指数增长速度快过人民平均收入的增长速度就一定不是好事，而一般平均工资的增长速度很难超越3%～4%。

居民消费价格指数是度量居民生活消费品和服务价格水平随着时间变动的相对数，综合反映居民购买的生活消费品和服务价格水平的变动情况。它是进行国民经济核算、宏观经济分析和预测、实施价格总水平调控的一项重要指标，并且世界各国一般用消费价格指数作为测定通货膨胀的主要指标。

一般来说当CPI同比增长大于3%时我们称通货膨胀；而当其大于5%时，我们称其为严重的通货膨胀。

(资料来源：http://baike.baidu.com/view/119777.htm? fromId=352674.)

2017年9月，全国居民消费价格同比上涨1.6%。其中，城市上涨1.7%，农村上涨1.4%；食品价格下降1.4%，非食品价格上涨2.4%；消费品价格上涨0.7%，服务价格上涨3.3%。1—9月平均，全国居民消费价格总水平比去年同期上涨1.5%。

9月，全国居民消费价格环比上涨0.5%。其中，城市上涨0.5%，农村上涨0.6%；食品价格上涨0.5%，非食品价格上涨0.5%；消费品价格上涨0.4%，服务价格上涨0.7%。

(1) 各类商品及服务价格同比变动情况。9月，食品烟酒价格同比下降0.4%，影响CPI下降约0.11个百分点。其中，畜肉类价格下降7.5%，影响CPI下降约0.37个百分点(猪肉价格下降12.4%，影响CPI下降约0.36个百分点)；鲜果价格下降3.0%，影响CPI下降约0.05个百分点；鲜菜价格下降1.0%，影响CPI下降约0.02个百分点；蛋类价格上涨3.7%，影响CPI上涨约0.02个百分点；水产品价格上涨2.5%，影响CPI上涨约0.05个百分点；粮食价格上涨1.7%，影响CPI上涨约0.03个百分点。

9 月，其他七大类价格同比均有不同程度的上涨。其中，医疗保健、居住、教育文化和娱乐价格分别上涨 7.6%、2.8%和 2.3%，生活用品及服务、其他用品和服务价格均上涨 1.4%，衣着、交通和通信价格分别上涨 1.3%和 0.5%。

(2) 各类商品及服务价格环比变动情况。9 月，食品烟酒价格环比上涨 0.4%。其中，蛋类价格上涨 5.3%，影响 CPI 上涨约 0.03 个百分点；鲜果价格上涨 1.9%，影响 CPI 上涨约 0.03 个百分点；畜肉类价格上涨 0.9%，影响 CPI 上涨约 0.04 个百分点(猪肉价格上涨 1.0%，影响 CPI 上涨约 0.03 个百分点)；水产品价格下降 1.8%，影响 CPI 下降约 0.03 个百分点。

9 月，其他七大类价格环比均有不同程度的上涨。其中，医疗保健、居住、交通和通信价格分别上涨 1.8%、0.3%和 0.1%，其他用品和服务、生活用品及服务价格分别上涨 0.2%和 0.1%；衣着、教育文化和娱乐价格均上涨 0.9%。

(资料来源：http://www.gov.cn/xinwen/2017-10/16/content_5232031.html.)

[B] 翘尾因素

“翘尾因素”或“翘尾影响”是指上年价格上涨(下降)对本年同比价格指数的滞后(延伸)影响。也就是在计算同比价格指数过程中，上年商品价格上涨(下降)对下一年价格指数的影响。

一般说来，上年价格上涨(下降)的时间早，则对下年指数的翘尾影响小；而上年价格上涨(下降)的时间晚，则对下年指数的翘尾影响大。上年调价幅度越大，时间越晚，翘尾影响就会愈加明显。因此在 CPI 的分析预测中，“翘尾因素”是一个不可忽视的问题。举个简单的例子：如某一商品 2016 年前 6 个月价格均为每千克 1 元，7 月上涨到 2 元，然后一直到 2017 年 12 月都保持在 2 元。那么，虽然 2017 年全年价格保持稳定，但如计算 2017 年前 6 个月与 2016 年前 6 月比的价格指数则为 200%，表明价格上涨 1 倍，这就是 2016 年 7 月价格上涨对下一年前 6 个月价格指数的滞后影响，简称“翘尾因素”。

(资料来源：http://www.stats.gov.cn/tjzs/zjCPI/t20110610_402731140.html.)

[C]价格水平与价格指数

价格水平是个绝对数，而价格指数是一个相对数。两者的用途不同：价格水平反映的是在一定时期商品合服务项目价格的水平，是一个静态指标。价格指数是用两个不同时期的价格水平相比较，反映一定时期内商品价格水平变动情况的统计指标，是一个动态指标。如果用两个不同时期的价格相比较，可以反映价格变动水平和变动程度。举个例子，今年某地大白菜 9 月每千克的平均价格为 2.8 元，8 月为 2.4 元，这说明了 8、9 月大白菜的“价格水平”。如果要反映大白菜价格 9 月比 8 月的变动程度，则需要计算价格指数。就是用 9 月大白菜的平均价格除以 8 月平均价格再乘以 100%求得，即 2.8/2.4×100%=116.7%，也就是说，9 月大白菜的价格比 8 月上涨了 16.7%。反过来说，如果今年 9 月大白菜的平均价格每千克为 2.4 元，8 月为 2.8 元，这说明 9 月大白菜的“价格水平”比 8 月下降了。9 月比 8 月大白菜的价格指数则为：2.4/2.8×100%=85.7%。也就是说，9 月比 8 月大白菜的价格下跌了 14.3%。依次类推，其他所有商品价格的变动程度，以及对整个商品价格总水平的影响大小，都可以通过价格指数来反映。

(来源：http://www.stats.gov.cn/tjzs/zjCPI/t20110610_402731128.html.)

第一节　时间序列概述

一、时间序列的概念

社会经济现象是处在不断变化发展过程中的，我们不仅要对社会经济现象的相互联系进行静态分析，还应该对社会经济现象在不同时间上的发展变化进行动态分析。时间序列分析就是运用时间序列的一系列动态指标，分析经济现象的变动过程和变动规律的一种动态分析方法。

时间序列，又称时间数列或动态数列，是指将某一指标在不同时间上的数值，按照时间(如年、季、月、日等)先后顺序排列而成的统计序列。如表 6-1 列举了我国 2010—2015 年国内生产总值、年末人口数全社会固定资产投资额、人均国内生产总值、居民消费价格总指数经济指标的时间序列。

表 6-1　我国 2010—2015 年重要经济指标

指标 年份	国内生产总值/亿元	年末人口数/万人	全社会固定资产投资额/亿元	人均国内生产总值生产总值/(元/人)	居民消费价格总指数/%(上年＝100)
2010	413 030.3	134 091	251 683.77	30 876	103.3
2011	489 300.6	134 735	311 485.13	36 403	105.4
2012	540 367.4	135 404	374 694.74	40 007	102.6
2013	594 244.4	136 072	446 294.09	43 852	102.6
2014	643 974.0	136 782	512 020.65	47 203	102.0
2015	676 708.0	137 462	562 000.00	49 351	101.4

资料来源：中华人民共和国国家统计局网站. http://www.stats.gov.cn/.

由表 6-1 可以看出，时间序列由两大要素构成；一是现象数值所属的时间，即现象发生的时间，可以表现为年、月、日或季、周等时间单位；二是统计指标的具体数值。

统计指标是用来说明某种社会经济现象在某一个方面的数量特征的，社会经济现象的发展变化必然在不同时间的指标数值上有所反映。时间序列就是反映该现象发展变化过程的历史记录。这对统计分析工作来说，具有十分重要的意义。首先，编制时间序列，可以描述社会经济现象的发展状态、发展趋势和结果；其次，通过对时间序列的分析，可以掌握社会经济现象发展变化的规律性；最后，时间序列可用来对社会经济现象的发展方向和速度进行预测。

二、时间序列分类

时间序列按其构成要素中统计指标值的表现形式，分为绝对数时间序列、相对数时间序列和平均数时间序列三种。其中绝对数时间序列是基本序列，相对数和平均数时间序列是在其基础上计算产生的，是它的派生序列。

(一) 绝对数时间序列

把一系列同类的总量指标按时间先后顺序排列而成的数列，称为绝对数时间序列，它

反映了现象在各期达到的绝对水平。例如,表 6-1 中的国内生产总值、年末人口数和全社会固定资产投资额都是绝对数时间序列。

按照总量指标反映的现象的时间状态不同,绝对数时间序列又分为时期序列和时点序列两类。

(1) 时期序列。当序列中排列的指标为时期指标,反映现象在各段时期内发展过程的总量时,就称为时期序列。例如表 6-1 中的我国 2010—2015 年国内生产总值序列就是一个时期序列。

时期序列的特点是:①序列中各个指标的数值是可以相加的,即相加具有一定的经济意义。由于时期序列中每个指标数值是表示现象在一段时期内发展变化的累计总量,所以相加后的数值就表示现象在更长一段时期内发展变化的累计总量。②序列中每一个指标数值的大小与所属的时间长短有直接的联系。在时期序列中,每个指标数值所对应的时间长度,称为“时期”。时期的长短,主要根据研究目的而定,可以是日、旬、月、季、年或更长时间。一般来说,时期越长,指标数值就越大,反之就越小。③序列中每个指标数值,通常是通过连续不断地登记而取得的。

(2) 时点序列。当序列中排列的指标为时点指标,反映现象在某一时点上所处的状态时,称该序列为时点序列。例如,表 6-1 中所列的我国 2010—2015 年全国年末人口数是时点序列。

时点序列有如下特点:①序列中各个指标数值是不能相加的,相加不具有实际经济意义。这是由于时点序列中每个指标数值都是表明某一时点上瞬间现象的总量,相加以后无法说明属于哪一时点上的数量。②序列中指标数值的大小与其间隔长短没有直接联系。在时点序列中,两个相邻的指标数值间相隔的时间距离称为“间隔”。由于时点序列每个指标数值只表明现象在某一时点上的状态数量,指标数值大小与时间间隔长短没有直接联系。例如,年末数值可能大于月末数值,也可能小于月末数值。③序列中指标的每个数值,通常都是间隔一定时期通过一次性登记取得的。

(二) 相对数时间序列

把一系列同类的相对指标按时间顺序排列而成的序列,称为相对数时间序列。它反映现象对比关系的发展变化情况,说明社会经济现象的比例关系、结构、速度的发展变化过程。如表 6-1 中的我国居民消费价格总指数就是相对数时间序列。在相对指标时间序列中,各个指标数值也是不能相加的。

(三) 平均数时间序列

把一系列同类的平均指标按时间顺序排列而成的序列,称为平均数时间序列。它反映现象一般水平的发展趋势。例如,表 6-1 中的人均国内生产总值即为平均数时间序列。在平均数时间序列中,各个指标数值相加也没有意义。

三、时间序列的编制原则

编制时间序列的目的是通过同一指标在不同时间上的数值对比来反映现象的发展过

程及其规律性。因此,保证序列中指标值之间的可比性,就成为编制时间序列应遵守的基本原则。具体来讲,应注意以下几点。

(一) 时期长短应尽量统一

在时期序列中,由于各指标数值的大小与时间的长短有直接的关系,所以,各个指标数值所属的时期长短应前后统一。时间越长,指标数值就越大,反之就越小。时期长短不一,往往就很难做直接比较。

对于时点序列来说,由于各个指标数值只反映现象在某一时点的状态,两时点间隔的长短,对时点指标数值大小没有直接影响,所以不存在时期长短应统一的问题。但为了更有利于对比,时点间隔最好能保持一致。

(二) 总体范围应该一致

所谓总体范围,即所研究的现象总体所包括的地区范围、隶属关系范围、行政区划范围等。在实际工作中,因为各时期行政区划、经济管理体制、基层单位的隶属关系等方面发生变化,统计口径往往前后不一致。比如,我们要研究某地区的人口变动情况,如果所研究地区的行政区划发生了变化,那么,变动前后的两个人口数值就不能直接对比,需要加以调整,然后再进行动态分析。

(三) 指标的经济内容应该一致

有时同一个指标虽然名称没有变动,但其经济内容却已有改变,这也是不可比的。对于这样的时间序列,根据不同时期指标数值的变化来进行分析,就会得出错误的结论。比如,我们编制某地 2000—2016 年增值税征收情况的时间序列,由于 2015 年进行营改增,使增值税的经济内容在 2015 年前后发生了变化,如果不首先加以调整,就不具有可比性,否则就会使所反映的问题失实。

(四) 计算口径应该统一

计算口径主要是指计算方法、计算价格以及计量单位等。比如,在研究某企业劳动生产率的增长情况时,如果各时期指标的计算方法不一致,有的按全部职工计算,有的按生产工人计算,指标数值的对比就失去意义;再比如,总产值有的按现价计算,有的按不变价格计算,或指标的计量单位有的按实物单位计量,有的按货币单位计量,这样所编制的时间序列势必会导致指标数值失去可比性。

四、时间序列的分解

由于社会经济现象的复杂性,时间序列各观察值的变化是由许多因素共同作用的结果,有些因素属于基本因素,对事物的发展起决定性作用,会使事物的发展呈现出一定的规律性;有些因素是偶然的非基本因素,对事物的发展只起局部的非决定性作用,使事物的发展表现出不规则的变动。为了研究社会经济现象发展变化的规律和趋势,并据此预测未来,就要将这些影响时间序列的因素加以分解并分别进行测定。在具体分析中可按

影响因素的性质不同加以分类，一般将社会经济现象时间序列的总变动分解为长期趋势、季节变动、循环变动和不规则变动四种主要因素。由于受这些因素中的一种因素或几种因素的共同影响，时间序列表现为不同的变动规律。

（一）长期趋势（T）

长期趋势是时间序列的主要构成要素，它是指现象在较长时期内持续发展变化的一种趋向或状态。可能呈现出不断向上增长的态势，也可能呈现为不断降低的趋势，是受某种固定的起根本性作用的因素影响的结果。例如，中国改革开放以来经济持续增长，表现为国内生产总值逐年增长的态势。

（二）季节变动（S）

本来意义上的季节变动是指受自然因素的影响，在一年中随季节的更替而发生的有规律的变动。现在对季节变动的概念有了扩展，对一年内或更短的时间内由于社会、政治、经济、自然因素影响，形成的以一定时期为周期的有规则的重复变动，都称为季节变动。例如，农业产品的生产、某些商品的销售量变动都呈现出季节性的周期变动。

（三）循环变动（C）

循环变动指某种现象在比较长的时期内呈现出的有一定规律性的周期性波动。循环变动与长期趋势不同，它不是单一方向的持续变动，而是有涨有落的交替波动。循环变动与季节变动也不同，循环变动的周期长短很不一致，不像季节变动那样有明显的按月或按季的固定周期规律，循环变动的规律性不甚明显，通常较难识别。

（四）不规则变动（I）

不规则变动指现象受众多偶然因素影响，而呈现的无规则的变动。包括由突发的自然灾害、意外事故或重大政治事件所引起的剧烈变动，也包括大量无可名状的随机因素干扰造成的起伏波动。是时间序列长期趋势、季节变动和循环变动后余下的变动。

时间序列的变动一般都是由于以上四种构成要素或其中一部分要素的影响而形成的。时间序列分析的任务之一，就是对时间序列中的这几种构成要素进行统计测定和分析，从中划分出各种要素的具体作用，揭示其变动的规律和特征，为认识和预测事物的发展提供依据。

进行时间序列分析的一个重要前提，就是了解四种变动因素：长期趋势 T、季节变动 S、循环变动 C 和不规则变动 I 以什么样的形式相结合（假设在时间数列中均包含有四种因素。当然，实际中并非如此）。把这四个影响因素同时间序列的关系用一定的数学关系式表示出来，就构成了时间序列的分解模型。将各影响因素分别从时间序列中分离出来并加以测定的过程，称为时间序列的构成分析。

按四种因素对时间序列的影响方式不同，时间序列可分解为多种模型，如乘法模型、加法模型、混合模型等。各种模型都是在一定的假定情况下成立的。其中最常用的是乘法模型。其形式如下：

$$Y = T \cdot S \cdot C \cdot I \tag{6-1}$$

式中,Y 为时间序列中的指标数值。

根据这个模型,要求出某个构成因素的影响,用其余构成部分除时间序列即可。例如,当求出长期趋势 T 以后,以 Y 除以 T,则可得不含长期趋势的派生时间序列:

$$\frac{Y}{T} = S \cdot C \cdot I$$

若再求出季节变动 S,用 S 去除,则可得不含长期趋势及季节变动的时间序列:

$$\frac{Y/T}{S} = C \cdot I$$

如果时间序列中仅含长期趋势和季节变动两个因素,则可以按以上相除的方法将两种因素分解开来分别进行分析。

第二节　时间序列的描述性分析

一、图形分析

在对时间序列进行分析时,可以首先利用时间序列数据做出一个散点图,然后根据图形观察数据随时间的变化模式及变化趋势。图形分析是分析时间序列形态的一种有效的分析方法,它对于时间序列的进一步分析和预测有很大的帮助。

二、水平分析

时间序列的水平分析是指利用一系列的水平指标对现象在某一时期或时点上发展变化的水平进行分析,以揭示社会经济现象的发展变化过程和规律。

(一) 发展水平分析

发展水平是时间序列中对应于具体时间的指标数值。反映处于量变过程中的事物在一定时期内或时点上所达到的规模或水平,简单地说,时间序列中,每个统计指标的数值就是发展水平。它一般是总量指标,如国内生产总值、轿车总产量等;也可以是相对指标,如计划完成程度、居民消费价格指数等;还可以是平均指标,如居民人均可支配收入、平均工资等。

如果用符号 $a_0, a_1, a_2, a_3, \cdots, a_{n-1}, a_n$ 代表时间序列中各个发展水平,通常,时间序列中第一个指标数值 a_0 叫作最初水平,最后一个指标数值 a_n 叫作最末水平,其余各个指标数值叫作中间水平。

此外,在时间序列分析中,常将所研究的那个时期的指标数值叫作报告期水平,用来进行比较的基础时期水平叫作基期水平。通常,报告期是指离分析者比较近的那个时期,基期是指离分析者较远的那个时期。报告期和基期的划分是相对的,而且是随着研究的问题不同而变化的。

(二) 平均发展水平分析

将不同时期的发展水平加以平均而得到的平均数叫作平均发展水平,又称为序时平

均数。它概括性地描述出现象在一段时期内所达到的一般水平。

序时平均数和静态平均数一样，都是将各个变量值的差异抽象化，反映一般水平。但两者也有显著的区别：静态平均数是将总体各单位在同一时间内的数量差异进行抽象，从时间截面上反映总体各单位标志值的一般水平；序时平均数则是将总体在不同时间内的数量差异进行抽象，从动态过程上反映总体指标值的一般水平。其次，两者的计算依据也不同，静态平均数是根据变量序列来计算的，而序时平均数是根据时间序列来计算的。

社会经济现象在不同时间上的发展变化总是不平衡的，为了概括地说明其发展变化的一般水平，反映其变动的规律性，就需要通过平均发展水平分析来解决。序时平均数可以解决时间序列中的某些可比性问题，便于对同一现象在不同历史阶段的变化状况进行比较，对不同单位、不同地区、不同部门或不同国家或地区在某一时间内某一现象发展的一般水平进行比较分析。

由于不同时间序列中观察值的表现形式不同，序时平均数有以下两种计算方法。

1. 根据绝对数时间序列计算序时平均数

绝对数时间序列序时平均数的计算方法是最基本的，它是计算相对数或平均数时间序列序时平均数的基础。由于绝对数时间序列有时期序列和时点序列之分，序时平均数的计算方法也有所不同。

1）时期序列的序时平均数

时期序列中的各观察值可以相加，形成一段时期内的累计总量，所以时期序列的序时平均数可直接用各时期的观察值之和除以时期项数来计算。其计算公式为

$$\bar{a}=\frac{a_1+a_2+\cdots+a_n}{n}=\frac{\sum_{i=1}^{n}a_i}{n} \tag{6-2}$$

式中，$\bar{a}$ 为序时平均数；a_i 为第 i 个时期的观察值；n 为观察值的个数(时期项数)。

在 Excel 中可以直接利用 AVERAGE()函数计算。

2）时点序列的序时平均数

根据所掌握的时点资料的不同，计算方法亦有所不同。时点数列一般都是不连续数列。但是，若是逐日记录而且逐日排列形成的时点数列资料则可将其看成连续的时点数列。以此为标准，可将时点数列分为连续时点数列和间断时点数列。

(1) 根据连续时点数列计算序时平均数。在连续时点数列中有间隔相等和间隔不等两种情况。

① 间隔相等的连续时点数列。即时点数列资料是逐日排列形成的。其计算公式为

$$\bar{a}=\frac{\sum a}{n} \tag{6-3}$$

例如，已知某单位一个月内每天当班的工人人数，要计算该月的每天平均工人数，可将每天的工人数相加，再除以该月的日历天数即可。在 Excel 中同样可以直接利用 AVERAGE()函数计算。

② 间隔不等的连续时点数列。如果被研究的现象每隔一段时间才有变动，则用每次

变动持续的间隔长度(f)为权数对各时点水平(a)加权,其计算公式为

$$\bar{a}=\frac{\sum af}{\sum f} \tag{6-4}$$

【例 6-1】 某企业 8 月 1—12 日设备数为 103 台。8 月 13 日到月底增加到 110 台,则该企业 8 月平均拥有设备数为

$$\bar{a}=\frac{\sum af}{\sum f}=\frac{103\times 12+110\times 19}{12+19}=\frac{3\ 326}{31}=107(\text{台})$$

(2) 根据间断时点数列计算序时平均数。间断时点数列中也有间隔相等和间隔不等两种情况。

① 间隔相等的间断时点数列。在实际统计工作中,对时点性质的指标,为了简化登记手续,往往间隔一定时间登记一次。如商业企业的商品库存、工业企业中的职工人数与流动资产等,都只统计月末数字,从而组成间隔相等的间断时点数列。一般情况下,这时可以假定所研究的现象在两个相邻时点之间的变动是均匀的,则相邻两个时点之间的序时平均数,可以通过两个相邻时点指标数值相加后除以 2 近似求得$\left(\bar{a}_i=\frac{a_{i-1}+a_i}{2}\right.$或$\frac{a_i+a_{i+1}}{2}$,即$\left.\frac{\text{期初}+\text{期末}}{2}\right)$,然后根据这些平均数,再用简单算术平均法求得整个研究时间的序时平均数$\left(\bar{a}=\frac{\sum_{i=1}^{n}\bar{a}_i}{n}\right)$。

【例 6-2】 已知某企业 2016 年 6—9 月各月末库存材料数如表 6-2 所示。计算第三季度平均库存材料数。

表 6-2 某企业 2016 年 6—9 月各月末库存材料数 吨

日　期	6 月 30 日	7 月 31 日	8 月 31 日	9 月 30 日
库存材料数	120	132	110	142

在 Excel 中的计算过程如下。

a. 计算月度平均。在 C3 单元格中输入公式"=(B2+C2)/2"即为 7 月份平均库存材料数;向右拖动鼠标复制公式即可计算出 8 月和 9 月的平均库存材料数。

b. 计算季度平均。在 B4 单元格中输入公式"=(C3+D3+E3)/3",即为第三季度的平均库存材料数。最终结果如图 6-1 所示。

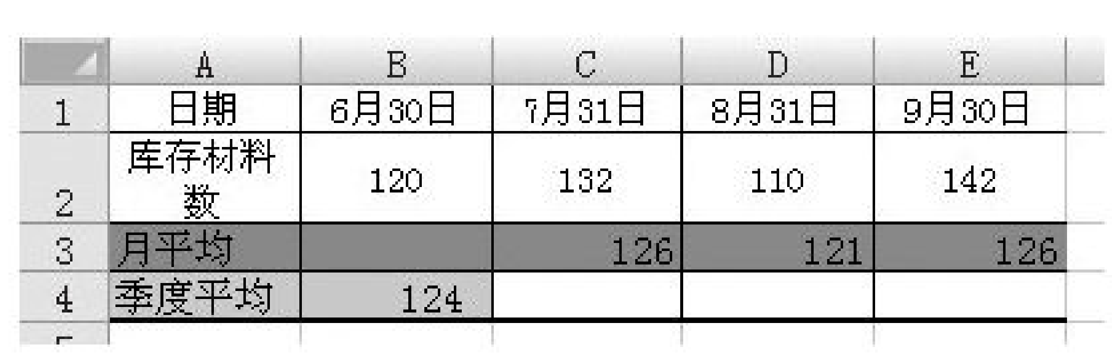


图 6-1 间隔相等的间断时点数列计算

此外,间隔相等的间断时点数列,计算序时平均数的公式还可表示为

$$\bar{a}=\frac{\frac{a_0}{2}+a_1+\cdots+a_{n-1}+\frac{a_n}{2}}{n} \tag{6-5}$$

式中,n 为时间项数。

② 间隔不等的间断时点数列。根据间隔不等的每期期末时点资料,则可用时间间隔

的长度为权数(f),对各相应时点的平均水平进行加权平均。其计算公式为

$$\bar{a}=\frac{\frac{a_0+a_1}{2}f_1+\frac{a_1+a_2}{2}f_2+\cdots+\frac{a_{n-1}+a_n}{2}f_n}{\sum f} \tag{6-6}$$

【例 6-3】 某银行储蓄存款余额数如表 6-3 所示。计算该行 2015 年储蓄存款的平均余额。

表 6-3　某银行 2015 年各时点的储蓄存款余额数　　万元

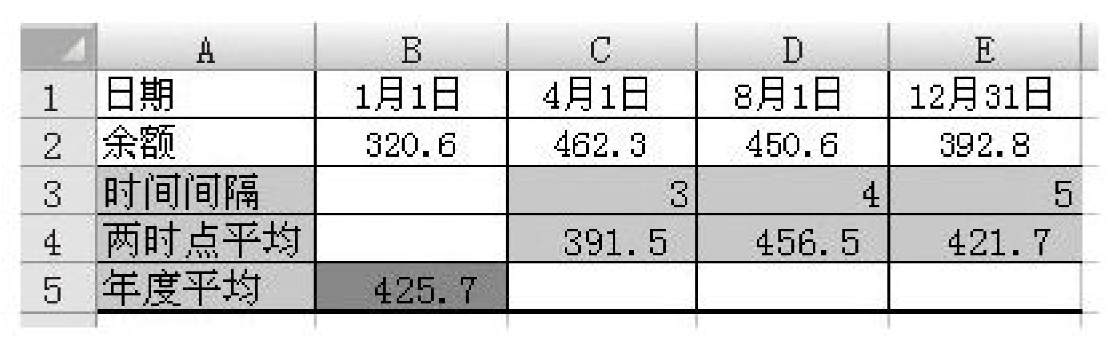
日 期	1 月 1 日	4 月 1 日	8 月 1 日	12 月 31 日
余额	320.6	462.3	450.6	392.8

在 Excel 中的计算过程如下。

a. 计算时间间隔,分别为:3、4、5。

b. 计算两间隔时点的平均数。在 C4 单元格中输入公式"=(B2+C2)/2",用鼠标向右拖动复制公式。

c. 计算年度的平均数。在 B5 单元格中输入公式"=(C4 * C3+D4 * D3+E4 * E3)/(C3+D3+E3)"。最终结果如图 6-2 所示。



	A	B	C	D	E
1	日期	1月1日	4月1日	8月1日	12月31日
2	余额	320.6	462.3	450.6	392.8
3	时间间隔		3	4	5
4	两时点平均		391.5	456.5	421.7
5	年度平均	425.7			

图 6-2　间隔不等的间断时点数列计算

根据间断时点数列计算序时平均数的前提是,假设被研究现象在相邻两个时点之间的变动是均匀的。当然这种现象的实际变动并不完全均匀,计算结果只能是实际值的近似值。同时,为了使计算结果能尽量反映实际情况,间断时点数列的间隔不宜过长。

2. **相对数或平均数时间序列的序时平均数**

相对数和平均数通常是由两个绝对数对比形成的,即观察值 $c_i=\frac{a_i}{b_i}$,计算序时平均数时,应先分别求出构成相对数或平均数的分子 a_i 和分母 b_i 的平均数,而后再进行对比,即得相对数或平均数序列的序时平均数。其基本公式为

$$\frac{\bar{a}}{\bar{b}}=\bar{c} \tag{6-7}$$

式中,$\bar{a}$ 和 $\bar{b}$ 可按绝对数时间序列序时平均数的计算方法求得。

【例 6-4】 某企业 2016 年各季度销售收入和流动资金资料如表 6-4 所示。

表 6-4　某企业 2016 年各季度销售收入与流动资金资料

时　　间	第一季度	第二季度	第三季度	第四季度
销售收入/万元	110	120	126	125
期初流动资金余额/万元	90	70	78	52
流动资金周转次数/次	1.375	1.62	2.1	1.64

又知,该年末流动资金余额为 100 万元。计算该企业 2016 年流动资金平均周转次数。

在 Excel 中的计算过程如下。

(1) 计算季度平均流动资金。在 B5 单元格中输入公式"=(B3+C3)/2",用鼠标向右拖动复制公式,即得各季度平均流动资金。

(2) 计算年度平均流动资金。在 B6 单元格中输入公式"=SUM(B5:E5)/4"。

(3) 计算年度平均销售收入。在 B7 单元格中输入公式"=SUM(B2:E2)/4"。

(4) 计算年度流动资金平均周转次数。在 B8 单元格中输入公式"=B7/B6"。最终结果如图 6-3 所示。

	A	B	C	D	E	F
1	时间	第一季度	第二季度	第三季度	第四季度	第四季度末
2	销售收入/万元	110	120	126	125	
3	期初流动资金余额/万元	90	70	78	52	100
4	流动资金周转次数/次	1.375	1.62	2.1	1.64	
5	季度平均流动资金	80	74	65	76	
6	年度平均流动资金	73.75				
7	年度平均销售收入	120.25				
8	年度流动资金平均周转次数	1.63				

图 6-3 相对数或平均数时间序列序时平均数的计算

三、增长量分析

(一) 增长量

增长量是时间序列中的报告期水平与基期水平之差,用于描述现象在观察期内增长的绝对量。若二者之差为正数,表示增长;若为负数,则表示下降。

由于采用的基期不同,增长量有逐期增长量和累计增长量之分。逐期增长量是报告期水平与前一时期水平之差,表示本期比前一时期增长的绝对数量;累计增长量是报告期水平与某一固定时期水平之差,说明报告期比某一固定时期增长的绝对数量。可用符号表示如下。

逐期增长量:

$$a_1 - a_0, a_2 - a_1, \cdots, a_n - a_{n-1} \tag{6-8}$$

累计增长量:

$$a_1 - a_0, a_2 - a_0, \cdots, a_n - a_0 \tag{6-9}$$

显然,两者之间具有一定的关系,累计增长量等于相应各个逐期增长量之和,即

$$a_n - a_0 = \sum_{i=1}^{n} (a_i - a_{i-1})$$

(二) 平均增长量

平均增长量是观察期各逐期增长量的平均数,用于描述在观察期内平均增长的数量。有两种计算方法。

第一种,根据逐期增长量求得,或根据累计增长量求得。其计算公式为

$$平均增长量 = \frac{逐期增长量之和}{逐期增长量个数} = \frac{累计增长量}{观察值个数 - 1}$$

$$=\frac{\sum_{i=1}^{n}(a_i-a_{i-1})}{n}=\frac{a_n-a_0}{n} \tag{6-10}$$

第二种，利用等差级数来计算。其计算公式为

$$平均增长量=\frac{2\left(\sum_{i=1}^{n}a_i-na_0\right)}{n(n+1)}$$

【例 6-5】 根据表 6-1 中我国国内生产总值资料，计算各年逐期增长量和各年以 2010 年为基期的累计增长量，年平均增长量。

在 Excel 中的计算步骤如下。

(1) 计算逐期增长量。在 C3 单元格中输入公式“＝B3－B2”，用鼠标向下拖动至 C7 单元格复制公式。

(2) 计算累计增长量。在 D3 单元格中输入公式“＝B3－$B$2”(注意 B2 单元格为绝对引用)，用鼠标向下拖动至 D7 单元格复制公式。

(3) 计算平均增长量。在 B12 单元格中输入公式“＝D7/5”，即得 5 年来的年平均增长量为 52 735.5 亿元。最终结果如图 6-4 所示。

	A	B	C	D
1	年份	国内生产总值/亿元	逐期增长量/亿元	累计增长量/亿元
2	2010	413 030.3	—	—
3	2011	489 300.6	76 270.3	76 270.3
4	2012	540 367.4	51 066.8	127 337.1
5	2013	595 244.4	54 877.0	182 214.1
6	2014	643 974.0	48 729.6	230 943.7
7	2015	676 708.0	32 734.0	263 677.7

图 6-4 增长量计算

请读者按第二种方法计算平均增长量。

四、速度分析

速度分析可以用于比较分析某种社会经济现象在不同发展阶段或不同地区、部门、国家之间的发展变化程度，也可以作为编制和检查国民经济计划的参考，是应用比较广泛的一种分析方法。

(一) 发展速度

发展速度是同一事物在两个不同时期发展水平对比的结果，用于描述现象在观察期内的相对发展变化程度，说明报告期水平是基期水平的百分之几或若干倍。当发展速度的计算结果大于 100%时，表明现象发展水平上升；反之，表明现象发展水平下降。当分子指标特别大而分母指标特别小的时候，也可以用倍数和翻番数表示；反之，可以用千分数或万分数表示。其计算公式为

$$发展速度=\frac{报告期水平}{基期水平}\times 100\% \tag{6-11}$$

由于采用的基期不同，发展速度可以分为环比发展速度和定基发展速度。环比发

展速度是报告期水平与前一期水平之比,说明现象逐期发展变化的程度;定基发展速度是报告期水平与某一固定时期水平之比,说明现象在整个观察期内总的发展变化程度。

设时间序列的观察值为 $a_i(i=1,2,\cdots,n)$,环比发展速度和定基发展速度的一般形式可以写为

$$\text{环比发展速度} = \frac{a_i}{a_{i-1}} \quad (i=1,2,\cdots,n) \tag{6-12}$$

$$\text{定基发展速度} = \frac{a_i}{a_0} \quad (i=1,2,\cdots,n) \tag{6-13}$$

环比发展速度与定基发展速度之间的关系如下。

(1) 观察期内各个环比发展速度的连乘积等于最末期的定基发展速度。

(2) 相邻两个定基发展速度的比值等于相应时期的环比发展速度。

即

$$\prod \frac{a_i}{a_{i-1}} = \frac{a_n}{a_0} \quad (\prod \text{为连乘符号}), \qquad \frac{a_i}{a_0} \div \frac{a_{i-1}}{a_0} = \frac{a_i}{a_{i-1}}$$

利用上述关系,可以根据一种发展速度去推算另一种发展速度。

(二) 增长速度

增长速度也称增长率,是增长量与基期水平之比,用于描述现象的相对增长速度。它可以根据增长量求得,也可以根据发展速度求得。其基本计算公式为

$$\text{增长速度} = \frac{\text{增长量}}{\text{基期水平}} = \frac{\text{报告期水平} - \text{基期水平}}{\text{基期水平}} = \text{发展速度} - 1 \tag{6-14}$$

发展速度与增长速度所说明的问题不同。发展速度说明现象在不同时期的发展程度,而增长速度则是说明其提高(或降低)的程度。因此,当发展速度大于 1 时,增长速度为正值,表示正增长;当发展速度小于 1 时,增长速度为负值,表示负增长。

由于采用的基期不同,增长速度也可分为环比增长速度和定基增长速度。前者是逐期增长量与前一期水平之比,用于描述现象逐期增长的程度;后者是累计增长量与某一固定时期水平之比,用于描述现象在观察期内总的增长程度。

环比增长速度和定基增长速度的公式可写为

$$\text{环比增长速度} = \frac{a_i - a_{i-1}}{a_{i-1}} = \frac{a_i}{a_{i-1}} - 1 \quad (i=1,2,\cdots,n) \tag{6-15}$$

$$\text{定基增长速度} = \frac{a_i - a_0}{a_0} = \frac{a_i}{a_0} - 1 \quad (i=1,2,\cdots,n) \tag{6-16}$$

需要指出,环比增长速度与定基增长速度之间没有直接的换算关系。在由环比增长速度推算定基增长速度时,可先将各环比增长速度加 1 后连乘,再将结果减 1,即得定基增长速度。

【例 6-6】 已知某企业 2011—2015 年的钢产量如表 6-5 所示。试计算各年的发展速度和增长速度。

表 6-5　某企业 2011—2015 年钢产量资料

年　　份	2011	2012	2013	2014	2015
钢产量/万吨	600	650	680	710	760

在 Excel 中的计算步骤如下。

(1) 计算环比发展速度。在 C3 单元格中输入公式“＝C2/B2 * 100”,用鼠标向右拖动复制公式。

(2) 计算定基发展速度。在 C4 单元格中输入公式“＝C2/$B$2 * 100”,用鼠标向右拖动复制公式。

(3) 计算环比增长速度。在 C5 单元格中输入公式“＝C3－100”,用鼠标向右拖动复制公式。

(4) 计算定基增长速度。在 C6 单元格中输入公式“＝C4－100”,用鼠标向右拖动复制公式。最终结果如图 6-5 所示。

	A		B	C	D	E	F
1	年　份		2011	2012	2013	2014	2015
2	钢产量/万吨		600	650	680	710	760
3	发展速度/%	环比		108.33	104.62	104.41	107.04
4		定基		108.33	113.33	118.33	126.67
5	增长速度/%	环比		8.33	4.62	4.41	7.04
6		定基		8.33	13.33	18.33	26.67

图 6-5　发展速度和增长速度计算

(三) 平均发展速度

平均发展速度是各个时期环比发展速度的平均数,用于描述现象整个观察期内平均发展变化的程度。平均发展速度可能大于 100%,也可能小于 100%,前者说明现象的发展水平是上升的,后者说明现象的发展水平是下降的。

由于环比发展速度是根据同一现象在不同时期发展水平对比而得到的动态相对数,它所依据的基数不同,因此,计算平均发展速度不能用算术平均数方法,而通常采用水平法和累计法。

1. 水平法(几何平均法)

水平法是根据各期的环比发展速度采用几何平均计算出来的,计算公式为

$$\overline{X}=\sqrt[n]{\frac{a_1}{a_0}\times\frac{a_2}{a_1}\times\cdots\times\frac{a_n}{a_{n-1}}}=\sqrt[n]{\prod\frac{a_i}{a_{i-1}}}=\sqrt[n]{\frac{a_n}{a_0}}\quad(i=1,2,\cdots,n)\tag{6-17}$$

式中,$\overline{X}$ 为平均发展速度;n 为环比发展速度的个数,它等于观察数据的个数减 1。

应用水平法计算平均发展速度的基本思想和原理是,从最初水平 a_0 出发,每期按平均发展速度 $\overline{X}$ 发展,经过 n 期后将达到最末水平 a_n,即 $a_0\times\overline{X}^n=a_n$。因此,用水平法计算的平均发展速度推算出的最末期的数值与最末期的实际观察值是一致的。从计算公式可以看出,按水平法计算的平均发展速度,实际上只与序列的最初观察值 a_0 和最末观察值 a_n 有关,而与其他各观察值无关,这一特点表明,水平法旨在考察现象在最末一期所应达到的发展水

平。所以,在实际应用中,如果我们所关心的是现象在最后一期应达到的水平例如最末期所达到的工业生产能力、产值、人口数的增长等,采用水平法计算平均发展速度比较合适。

2. **累计法(方程法)**

在一个时间序列中,各期实际水平之和为

$$a_1 + a_2 + \cdots + a_n = \sum_{i=1}^{n} a_i$$

在最初水平 a_0 的基础上,若各期的发展速度为 $X_i(i=1,2,\cdots,n)$,那么 $\sum_{i=1}^{n} a_i$ 也可以表示为

$$a_0 X_1 + a_0 X_1 X_2 + \cdots + a_0 X_1 X_2 \cdots X_n = \sum_{i=1}^{n} a_i$$

如果在最初水平 a_0 的基础上,每一期均按固定的平均发展速度 $\overline{X}$ 去发展,各期按固定平均发展速度计算的水平分别为 $a_0 \overline{X}^i(i=1,2,\cdots,n)$,并设定计算期的各期发展水平之和等于各期实际水平总和 $\sum_{i=1}^{n} a_i$,那么

$$a_0 \overline{X} + a_0 \overline{X}^2 + \cdots + a_0 \overline{X}^n = \sum_{i=1}^{n} a_i$$

将两边同除以 a_0,则有

$$\overline{X} + \overline{X}^2 + \cdots + \overline{X}^n = \frac{\sum_{i=1}^{n} a_i}{a_0}$$

解此高次方程所得 $\overline{X}$ 的正根,就是按累计法所求得的平均发展速度。

应用累计法计算平均发展速度的特点,是着眼于各期发展水平的累计之和。若在实际中侧重于考察现象各期发展水平的总和,如累计新增固定资产数、累计毕业生人数等,则应采用累计法比较合适。

(四) 平均增长速度

平均增长速度(平均增长率)则是用于描述现象在整个观察期内平均增长变化的程度,它通常用平均发展速度减 1 来求得。平均增长速度是现象各期环比增长速度的动态平均数,它表明现象在一个较长时期内平均增长的水平。平均增长速度与平均发展速度有密切的联系:

$$\text{平均增长速度} = \text{平均发展速度} - 1 \tag{6-18}$$

平均增长速度可能是正数,也可能是负数,当平均发展速度大于 1 时,平均增长速度为正数,说明现象在一段时期内发展水平增长的平均程度;当平均发展水平小于 1 时,平均增长速度是负数,说明现象在一段时期内发展水平降低的平均程度。

在进行增长分析时,需要注意的是,在有些情况下,不能单纯就速度论速度,要注意把速度分析与水平分析相结合。

【例 6-7】 假定有两个生产条件基本相同的企业,报告期与基期的利润额及有关速度资料如表 6-6 所示。

表 6-6　甲、乙两企业的有关资料

时　间	甲　企　业		乙　企　业	
	利润额/万元	增长率/%	利润额/万元	增长率/%
报告期	1 100	10	15	50
基　期	1 000	—	10	—

如果不看利润额的绝对值，仅就速度对甲、乙两个企业进行分析评价，可以看出乙企业的利润增长速度比甲企业高出 4 倍。如果就此得出乙企业的生产经营业绩比甲企业要好得多，这样的结论就是不切实际的。因为速度是一个相对数，它与对比的基期值的大小有很大的关系。大的速度背后，其隐含的增长绝对值也可能很小；小的速度背后，其隐含的增长绝对值也可能很大。这就是说，由于对比的基点不同，可能会造成速度数值上的较大差异，进而造成速度上的虚假现象。上述例子表明，由于两个企业的生产起点不同，基期的利润额不同，才造成二者速度上的较大差异。从利润的绝对额来看，两个企业的速度每增长一个百分点所增加的利润绝对额是不同的。在这种情况下，我们需要将速度和绝对水平结合起来进行分析，通常要计算增长 1%的绝对值来弥补速度分析中的局限性。

增长 1%的绝对值表示速度每增长一个百分点而增加的绝对数量，其计算公式为

$$\text{增长 }1\%\text{ 绝对值} = \frac{\text{逐期增长量}}{\text{环比增长速度} \times 100} = \frac{\text{前期水平}}{100} \tag{6-19}$$

根据表 6-6 的资料计算，甲企业速度每增长 1%的利润额为 10 万元，而乙企业则为 0.1 万元，甲企业远高于乙企业。这说明甲企业的生产经营业绩不是比乙企业差，而是更好。

第三节　时间序列的长期趋势分析

认识和了解社会经济现象的长期趋势，使之能够把握客观现象的基本特点和长期趋势，预测现象的发展趋势。通过长期趋势的测定，可以剔除时间序列中的季节变动、循环变动和不规则变动等因素的影响，有利于我们了解现象的发展过程和发展前景，并以此作为预测未来的依据。

一、线性趋势

线性趋势也称直线趋势，是指现象随着时间的推移而呈现出稳定增长或下降的线性变化规律。线性趋势的分析方法有很多，这里只介绍常用的几种。

（一）移动平均法

移动平均法是趋势变动分析的一种较简单的常用方法。当时间序列的变动趋势为线性状态时，可采用简单移动平均法进行描述和分析。该方法的基本思想和原理是，通过扩大原时间序列的时间间隔，并按一定的间隔长度逐期移动，分别计算出一系列移动平均数，由这些平均数形成的新的时间序列对原时间序列的波动起到一定的修匀作用，削弱了

原序列中短期偶然因素的影响,从而呈现出现象发展的变动趋势。

设观测的时间序列为 $y_1,y_2,\cdots,y_n$,则 $k(1<k<n)$期的一次移动平均的计算公式为

$$M_t^{(1)}=\frac{1}{k}(y_t+y_{t-1}+\cdots+y_{t-k+1})\quad(t=1,2,\cdots,n)\tag{6-20}$$

移动平均法只使用最近 K 期的数据,在每次计算移动平均值时,移动的时间间隔均为 K。此方法也适合于对只含随机波动的时间序列进行分析和预测。运用此方法的关键是确定合理的移动时间间隔。对于不同的时间序列,采用不同的移动间隔 K 对于时间序列的长期趋势反映准确度是不同的。选择间隔 K 时,可以采用试验的办法,选择一个使均方误差达到最小的间隔 K。

【例 6-8】 根据表 4-14 中的国内生产总值数据,分别取移动间隔 $K=3$ 和 $K=5$,在 Excel 中进行分析。

在 Excel 2013 中的操作步骤如下。

(1) 单击“数据”菜单下的“数据分析”按钮,在弹出的“数据分析”对话框中选择“移动平均”,如图 6-6 所示,单击“确定”按钮。

(2) 在弹出的移动平均设置对话框中,在输入区域内选择国内生产总值数据区域,设置间隔为 3,即 3 期移动平均,输出区域设置为 C2,如图 6-7 所示。



图 6-6 “数据分析”对话框

(3) 单击“确定”按钮,即可得到 3 期移动平均的数值。

(4) 用同样的方法计算 5 期移动平均的数值,最终结果如图 6-8 所示。



图 6-7 移动平均设置



	A	B	C	D
1	年份	国内生产总值	三期移动平均	五期移动平均
2	2010	413 030.3	#N/A	#N/A
3	2011	489 300.6	#N/A	#N/A
4	2012	540 367.4	480 899.4	#N/A
5	2013	595 244.4	541 637.5	#N/A
6	2014	643 974.0	593 195.3	536 383.3
7	2015	676 708.0	638 642.1	589 118.9

图 6-8 Excel 计算的移动平均值

(二) 指数平滑法

指数平滑法是另一种计算长期变动趋势的方法。移动平均法在逐期进行移动平均时,将每个样本点的作用等同对待。但在时间序列中,越靠近当前时刻的观察值越能反映当前时刻的性质,而远离当前时刻的观察值对当前时刻的代表性越弱。指数平滑法在计算移动平均时引入一个权数,使离当前时刻越近的样本点所起的作用越大。该方法使得第$(t+1)$期的预测值等于第 t 期的实际值与第 t 期的预测值的加权平均值。指数平滑法有一次指数平滑、二次指数平滑、三次指数平滑等,这里我们主要介绍一次指数平滑。

一次指数平滑是以一段时期的预测值与实际值的线性组合作为第$(t+1)$期的预测值，其公式为

$$F_{t+1}=\alpha X_t+(1-\alpha)F_t \tag{6-21}$$

式中，X_t 为时间序列第 t 时期的实际值；F_t 为第 t 期的预测值；α 为平滑系数，在 Excel 中$(1-\alpha)$称为阻尼系数，是介于 0 和 1 之间的数。

系数 α 的大小决定了平滑的程度，它与移动平均的间隔有类似的性质，适当选取 α 值是决定指数平滑结果优劣的重要因素。一般通过多次试算，然后比较各种 α 的趋势线以选出一个最优值。

需要注意的是，以指数平滑预测的结果存在滞后偏差，即当时间序列呈下降趋势时，预测值往往偏高；反之，则偏低。另外，一次指数平滑预测只能做下一期的预测。

【例 6-9】 根据表 4-14 中的国内生产总值数据，分别取移动平滑系数 $\alpha=0.3$ 和 $\alpha=0.8$，在 Excel 中进行分析。

在 Excel2013 中的操作步骤如下。

（1）单击“数据”菜单下的“数据分析”按钮，在弹出的“数据分析”对话框中选择“指数平滑”选项，再单击“确定”按钮。

（2）在弹出的“指数平滑”对话框中，在输入区域内选择城镇居民人均可支配收入数据区域，阻尼系数为“$1-\alpha=0.7$”，输出区域设置为 C2，如图 6-9 所示。

（3）单击“确定”按钮，即可得到平滑系数为 0.3 的预测值。

（4）用同样的方法计算 $\alpha=0.8$ 的预测数值，最终结果如图 6-10 所示。



图 6-9　指数平滑设置

	A	B	C	D
1	年份	国内生产总值	α=0.3	α=0.8
2	2010	413 030.3	#N/A	#N/A
3	2011	489 300.6	413 030.3	413 030.3
4	2012	540 367.4	435 911.4	474 046.5
5	2013	595 244.4	467 248.2	527 103.2
6	2014	643 974.0	505 647.1	581 616.2
7	2015	676 708.0	547 145.1	631 502.4

图 6-10　指数平滑结果

（三）直线趋势方程拟合法

直线趋势方程拟合法是利用直线回归的方法对原时间序列拟合线性方程，消除其他成分变动，从而揭示出数列长期直线趋势的方法。当现象的逐期增长量(一次差)大体相同时，可以考虑拟合直线趋势方程。直线趋势方程的一般形式为

$$\hat{Y}_t=a+b\cdot t \tag{6-22}$$

式中，$\hat{Y}_t$ 为时间数列 Y_t 趋势值；t 为时间；a 为截距项，是 $t=0$ 时 $\hat{Y}_t$ 的初始值；b 为趋势线

斜率,表示时间 t 变动一个单位时趋势值 $\hat{Y}_t$ 的平均变动数量。

趋势方程中的两个未知常数 a 和 b 通常按最小二乘法求得。该方法是根据回归分析中的最小二乘法原理,对时间序列配合一条趋势线,使之满足条件：各实际观察值(Y_t)与趋势值($\hat{Y}_t$)的离差平方和为最小,即 $\sum(Y_t-\hat{Y}_t)^2=$ 最小值。然后,根据所确定的趋势线计算出各个时期的趋势值,从而观察和描述现象发展的变动趋势,并对未来的趋势值进行预测。

将 $\hat{Y}_t=a+b\cdot t$ 代入 $\sum(Y_t-\hat{Y}_t)^2=$ 最小值,得 $\sum(Y_t-a-b\cdot t)^2=$ 最小值,使用数学中求函数极值的方法可得出一对联立方程

$$\begin{cases}\sum y-na-b\sum t=0\\ \sum ty-a\sum t-b\sum t^2=0\end{cases}\tag{6-23}$$

求出参数 a、b 的公式为

$$\begin{cases}b=\dfrac{n\sum tY-\sum t\sum Y}{n\sum t^2-\left(\sum t\right)^2}\\ a=\bar{Y}-b\cdot\bar{b}\end{cases}\tag{6-24}$$

把求出的 a、b 值代入 $\hat{Y}_t=a+b\cdot t$,即可得到直线趋势方程。

在 Excel 中,可以用数据分析工具库计算直线趋势方程的系数 a、b 及其相关统计量。具体操作步骤如下：

(1) 单击“数据”→“数据分析”→“回归”,弹出“回归”对话框,设置相关区域,如图 6-11 所示。



图 6-11 回归对话框

(2) 单击“确定”按钮,即可得到输出结果,如表 6-7 所示。其中：$a=112\ 731.792\ 7$,

b=52 305.198 18。

表 6-7 2005—2015 年我国国内生产总值数据及回归统计结果

年 份	t	国内生产总值	年 份	t	国内生产总值
2005	1	183 867.9	2011	7	489 300.6
2006	2	210 871.0	2012	8	540 367.4
2007	3	270 232.3	2013	9	595 244.4
2008	4	319 515.5	2014	10	643 974.0
2009	5	349 081.4	2015	11	676 708.0
2010	6	413 030.3			

SUMMARY OUTPUT

回归统计	
Multiple R	0.997 108 125
R Square	0.994 224 613
Adjusted R Square	0.993 582 903
标准误差	13 936.965 5
观测值	11

方差分析

	df	SS	MS	F	Significance F	
回归分析	1	3.00942E+11	3.01E+11	1549.337	2.19242E−11	
残差	9	1 748 151 067	1.94E+08			
总计	10	3.0269E+11				
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	112 731.792 7	9 012.614 566	12.508 22	5.4E−07	92 343.842 13	133 119.7
X Variable 1	52 305.198 18	1 328.837 522	39.361 62	2.19E−11	49 299.158 86	55 311.24

二、非线性趋势

当现象的长期趋势不是线性的，但又有一定的规律性，这时称现象的长期趋势为非线性趋势。若现象呈现出某种非线性状态，就需要配合适当的趋势曲线。趋势曲线的形式很多，有抛物线型、指数曲线型、修正指数曲线型、龚珀茨曲线型、Logistic 曲线型等。下面介绍几种常见曲线的拟合方法。

（一）二次抛物曲线

二次抛物曲线的标准形式为

$$\hat{Y}_t = a + bt + ct^2 \tag{6-25}$$

该曲线的特点是各期观察值的二次差相等。所谓二次差是各观察值逐期增长量的逐期增

长量。

曲线中的三个未知参数 a、b、c,同样可根据最小二乘法求得。即 $\sum(Y_t-\hat{Y}_t)^2=$ 最小值,把 $\hat{Y}_t=a+bt+ct^2$ 代入,得 $\sum(Y_t-a-bt-ct^2)^2=$ 最小值,利用函数极值法导出下列三个标准求解方程式

$$\begin{cases}\sum Y=na+b\sum t+c\sum t^2\\ \sum tY=a\sum t+b\sum t^2+c\sum t^3\\ \sum t^2Y=a\sum t^2+b\sum t^3+c\sum t^4\end{cases}\tag{6-26}$$

解此三元一次方程组,把求出的 a、b、c 的值代入 $\hat{Y}_t=a+bt+ct^2$,即可得到二次抛物线曲线趋势方程。

在手工计算中,为了简便起见,通常是根据 n 的奇偶性将 t 进行线性变换,使得 $\sum t=0$,$\sum t^3=0$,把式(6-26)简化为式(6-27)再进行求解。

$$\begin{cases}\sum Y=na+c\sum t^2\\ \sum tY=b\sum t^2\\ \sum t^2Y=a\sum t^2+c\sum t^4\end{cases}\tag{6-27}$$

但是,利用 Excel 进行求解时,不必多此一举,直接利用 LINEST()函数求解即可。

【例 6-10】 根据图 6-12 中的金属切削机床产量数据,确定二次抛物线曲线趋势方程,并利用趋势方程计算序列各期的预测值。

在 Excel 2013 中的操作步骤及结果如图 6-12 所示。

	A	B	C	D	E	F	G
1	年份	t	t^2	金属切削机床产量/万台	预测值		
2	2000	1	1	17.7	15.70		
3	2001	2	4	25.6	23.68		
4	2002	3	9	30.9	31.00		
5	2003	4	16	30.6	37.67		
6	2004	5	25	38.9	43.67		
7	2005	6	36	51.1	49.02		
8	2006	7	49	57.3	53.70		
9	2007	8	64	64.7	57.73		
10	2008	9	81	61.7	61.10		
11	2009	10	100	58.6	63.82		
12	a=	7.065		b=	8.97	c=	-0.329
13	趋势方程	Y=7.065+8.97t-0.329t^2					

图 6-12 Excel 中计算二次抛物线曲线趋势及预测结果

图 6-12 中,首先在 B12 中输入公式"=INDEX(LINEST(D2:D11,B2:C11,1,1),1,3)",即得 a 的值;在 E12 中输入公式"=INDEX(LINEST(D2:D11,B2:C11,1,1),1,2)",即得 b 的值;在 G12 中输入公式"=INDEX(LINEST(D2:D11,B2:C11,1,1),1,1)"即得 c 的值;然后在 E2 中输入公式"=$B$12+$E$12 * B2+$G$12 * C2",即得 2000 年的预测值,向下拖动鼠标复制公式即可得到历年的预测值。

（二）指数曲线

当现象的长期趋势每期大体按相同的增长速度递增或递减变化时，可拟合如下指数曲线方程：

$$\hat{Y}_t = ab^t \tag{6-28}$$

指数曲线的特点是各期的环比增长速度相同，或者时间序列的逐期趋势值按一定的百分比递增或衰减。

为估计 a、b，在手工计算中，一般首先将式(6-28)两端取对数，得

$$\lg\hat{Y}_t = \lg a + t\lg b$$

然后运用最小二乘法可以得到标准方程：

$$\begin{cases} \sum \lg Y = n\lg a + \left(\sum t\right)\lg b \\ \sum t\lg Y = \left(\sum t\right)\lg a + \left(\sum t^2\right)\lg b \end{cases} \tag{6-29}$$

估计出 $\lg a$ 和 $\lg b$，再取反对数，即可得参数 a、b 的估计值。

但在 Excel 中，可以无须事先做任何变换处理，直接利用 LOGEST()函数计算 a、b 的估计值以及相关统计量的值。

【例 6-11】 根据图 6-11 中的轿车产量数据，确定指数曲线趋势方程，并利用趋势方程计算序列各期的预测值。

在 Excel 2013 中的操作步骤及结果如图 6-13 所示。

	A	B	C	D
1	年份	t	轿车产量/万台	预测值
2	2000	1	60.7	65.88
3	2001	2	70.36	87.01
4	2002	3	109.2	114.93
5	2003	4	202.01	151.81
6	2004	5	231.4	200.53
7	2005	6	277.01	264.87
8	2006	7	386.94	349.86
9	2007	8	479.78	462.12
10	2008	9	503.74	610.41
11	2009	10	748.48	806.27
12	a=	49.87	b=	1.32
13	趋势方程	$\hat{Y}_t=49.87*1.32^t$		

图 6-13　Excel 中计算指数曲线趋势及预测结果

（三）修正指数曲线

在一般指数曲线的基础上增加一个常数 K，即为修正指数曲线(modified exponential curve)。其一般形式为

$$\hat{Y}_t = K + ab^t \tag{6-30}$$

式中，K，a，b 为未知参数，$K>0$，$a\neq 0$，$0<b\neq 1$。

修正指数曲线用于描述这样一类现象：初期增长迅速，随后增长率逐渐降低，最终则以 K 为增长极限。即当 $K>0$，$a<0$，$0<b<1$ 时，$t\to\infty$，$\hat{Y}_t\to K$。例如，某种刚问世的新产品，初期销售量增长可能很快，当市场拥有量接近饱和时，销售量逐渐趋于某一稳定的水平。现实生活中有许多事物的发展过程符合修正指数曲线的形式。

修正指数曲线中的三个未知参数 K、a、b 可用三和法求解。其基本思想是：把整个时间序列分成相等项数的三个组，每个组有 m 项，根据趋势值($\hat{Y}_t$)的三个局部总和分别等于原数列观察值(Y_t)的三个局部总和来确定三个参数。

设观察值的三个局部总和分别为 S_1、S_2、S_3，即

$$S_1 = \sum_{t=0}^{m-1} Y_t,\quad S_1 = \sum_{t=m}^{2m-1} Y_t,\quad S_3 = \sum_{t=2m}^{3m-1} Y_t \tag{6-31}$$

根据三和法的要求,得

$$\begin{cases} S_1 = mK + a + ab + ab^2 + \cdots + ab^{m-1} \\ \quad = mK + a(1 + b + b^2 + \cdots + b^{m-1}) \\ S_2 = mK + ab^m + ab^{m+1} + \cdots + ab^{2m-1} \\ \quad = mK + ab^m(1 + b + b^2 + \cdots + b^{m-1}) \\ S_3 = mK + ab^{2m} + ab^{2m+1} + \cdots + ab^{3m-1} \\ \quad = mK + ab^{2m}(1 + b + b^2 + \cdots + b^{m-1}) \end{cases} \tag{6-32}$$

由式(6-32),解得

$$\begin{cases} b = \left(\dfrac{S_3 - S_2}{S_2 - S_1}\right)^{\frac{1}{m}} \\ a = (S_2 - S_1)\dfrac{b-1}{(b^m-1)^2} \\ K = \dfrac{1}{m}\left(S_1 - \dfrac{a(b^m-1)}{b-1}\right) \end{cases} \tag{6-33}$$

【例 6-12】 我国 1995—2009 年成品糖的产量数据如图 6-14 所示。试确定修正指数曲线趋势方程并利用其计算序列各期的预测值。

	A	B	C	D	E	F
1	年份	t	成品糖产量/万吨	预测值		
2	1995	1	558.64	674.20		
3	1996	2	640.20	693.24		
4	1997	3	702.58	714.81		
5	1998	4	826.00	739.25		
6	1999	5	861.00	766.92		
7	**S1**	—	**3 588.42**	**3 588.42**		
8	2000	6	700.00	798.27		
9	2001	7	653.10	833.78		
10	2002	8	926.00	874.00		
11	2003	9	1083.94	919.55		
12	2004	10	1033.70	971.15		
13	**S2**	—	**4 396.74**	**4 396.74**		
14	2005	11	912.37	1 029.59		
15	2006	12	949.07	1 095.79		
16	2007	13	1 271.38	1 170.78		
17	2008	14	1 432.61	1 255.71		
18	2009	15	1 338.35	1 351.91		
19	**S3**	—	**5 904**	**5 903.78**		
20	b=	1.13	a=	126.72	K=	530.66
21	趋势方程	$y=530.66+126.72\times1.13^t$				

图 6-14 Excel 中计算修正指数曲线趋势及预测结果

在 Excel 2013 中的操作步骤及结果如图 6-14 所示。

首先在 C7 单元格中输入公式"=SUM(C2:C6)",在 C13 单元格中输入公式"=SUM(C8:C12)",在 C19 单元格中输入公式"=SUM(C14:C18)";然后在 B20 单元格中输入公式"=((C19-C13)/(C13-C7))^(1/5)",即为参数 b 的值,在 D20 单元格中输入公式"=(C13-C7)*(B20-1)/(B20*(B20^5-1)^2)",即为参数 a 的值,在 F20 单元

格中输入公式“=(1/5)*(C7-(D20*B20*(B20^5-1)/(B20-1)))”,即为常数 K 的值;最后在单元格 E2 中输入公式“=\$F\$20+\$D\$20*\$B\$20^B2”,用鼠标向下拖动复制公式即可,注意 E7、E13、E19 单元格不参与公式复制。

(四) 龚珀茨曲线

龚珀茨曲线(Gompertz curve)是以英国统计学家和数学家 B. Gompertz 的名字而命名的。其曲线方程为

$$\hat{Y}_t = Ka^{b^t} \tag{6-34}$$

式中,K,a,b 为未知参数,$K>0,0<a\neq1,0<b\neq1$

Gompertz 曲线所描述的现象的特点是:初期增长缓慢,以后逐渐加快,当达到一定程度后,增长率又逐渐下降,最后接近一条水平线。该曲线的两端都有渐进线,其上渐进线为 $Y=K$,下渐进线为 $Y=0$。该曲线多用于新产品的研制、发展、成熟和衰退分析,工业生产的增长、产品的寿命周期、一定时期内人口增长等现象也符合该曲线。

为求曲线中的参数,将式(6-34)两边取对数,可得

$$\lg\hat{Y}_t = \lg K + (\lg a)b^t \tag{6-35}$$

然后仿照修正指数曲线的参数求法,求出 $\lg a$、$\lg K$、b,取 $\lg a$ 和 $\lg K$ 的反对数求得 a 和 K。设

$$S_1 = \sum_{t=0}^{m-1}\lg Y_t,\quad S_2 = \sum_{t=m}^{2m-1}\lg Y_t,\quad S_3 = \sum_{t=2m}^{3m-1}\lg Y_t \tag{6-36}$$

则有

$$\begin{cases} b = \left(\dfrac{S_3 - S_2}{S_2 - S_1}\right)^{\frac{1}{m}} \\ \log a = (S_2 - S_1)\dfrac{b-1}{(b^m-1)^2} \\ \log k = \dfrac{1}{m}\left(S_1 - \dfrac{(b^m-1)}{b-1}\right)\log a \end{cases} \tag{6-37}$$

【例 6-13】 根据图 6-15 所示对 2003—2017 年成品糖的产量数据,确定 Gompertz 曲线趋势方程并利用其计算序列各期的预测值。

在 Excel 2013 中的操作步骤及结果如图 6-15 所示。

首先,把 Y 的值转换为 $\lg Y$ 的值,在 D2 单元格中输入公式“=log(C2)”,用鼠标向下拖动复制公式即可,注意 D7、D13、D19 单元格不参与公式复制。

其次,在 C7 单元格中输入公式“=SUM(C2:C6)”,在 C13 单元格中输入公式“=SUM(C8:C12)”,在 C19 单元格中输入公式“=SUM(C14:C18)”,用鼠标向右拖曳至 E7 单元格复制公式。

再次,在 B20 单元格中输入公式“=((D19-D13)/(D13-D7))^(1/5)”,即为参数 b 的值,在 D20 单元格中输入公式“=(D13-D7)*(B20-1)/(B20*(B20^5-1)^2)”,即为 $\lg a$ 的值,在 F20 单元格中输入公式“=(1/5)*(D7-(B20*(B20^5-1)/(B20-1)

	A	B	C	D	E	F
1	年份	t	成品糖产量/万吨	lg（Y）	预测值	
2	2003	1	558.64	2.75	667.45	
3	2004	2	640.20	2.81	685.99	
4	2005	3	702.58	2.85	706.76	
5	2006	4	826.00	2.92	730.09	
6	2007	5	861.00	2.94	756.37	
7	**S1**	—	**3 588.42**	**14.25**	**3 546.67**	
8	2008	6	700.00	2.85	786.05	
9	2009	7	653.10	2.81	819.68	
10	2010	8	926.00	2.97	857.94	
11	2011	9	1 083.94	3.04	901.63	
12	2012	10	1 033.70	3.01	951.72	
13	**S2**	—	**4 396.74**	**14.68**	**4 317.02**	
14	2013	11	912.37	2.96	1 009.42	
15	2014	12	949.07	2.98	1 076.22	
16	2015	13	1 271.38	3.10	1 153.98	
17	2016	14	1 432.61	3.16	1 245.04	
18	2017	15	1 338.35	3.13	1 352.36	
19	**S3**	—	**5904**	**15.32**	**5 837.01**	
20	b=	1.09	lg$a$=	0.12	lg$K$=	2.69
21			a=	1.33	k=	490.03
22	趋势方程	y=490.03*1.33$^{1.09^t}$				

图 6-15 Excel 中计算 Gompertz 曲线趋势及预测结果

＊D20))”，即为常数 lg$K$ 的值；在单元格 D21 中输入公式“＝10^D20”，即为 a 的值，在 F21 中输入公式“＝10^F20”，即为 K 的值。

最后，在 E2 中输入公式“＝$F$21＊($D$21^($B$20^B2))”，用鼠标向下拖动复制公式即可，注意 E7、E13、E19 单元格不参与公式复制。

三、趋势线的选择

趋势模型的选择是定性分析和定量分析相结合的过程。

首先，进行定性分析。应了解所研究现象的客观性质及其相关的理论知识，根据现象观察值的发展变化规律及其散点图的形态确定适当的趋势线类型。这在一定程度上取决于研究者的个人经验及理论知识水平。

其次，可根据所观察时间序列的数据特征，按以下标准考虑选择趋势线。

(1) 若观察值的一次差(逐期增长量)大致相同，可配合直线。

(2) 若二次差(逐期增长量的逐期增长量)大致相同，可配合二次曲线。

(3) 若各观察值的环比增长速度大致相同，可配合指数曲线。

(4) 若各观察值一次差的环比速度大致相同，可配合修正指数曲线；

(5) 若各观察值对数一次差的环比速度大致相同，可配合 Gompertz 曲线。

(6) 若各观察值倒数一次差的环比速度大致相同，可配合逻辑斯蒂曲线。

最后，如果对同一时间序列有几种趋势线可供选择，可通过参数估计的若干统计量指标比较选择。

第四节 时间序列的季节变动分析

一、季节变动的含义

季节变动是指客观现象因受自然因素或社会经济因素影响,在一年内形成的有规律的周期性变动。它是时间序列的又一个主要构成要素。季节变动在现实生活中经常遇到,如商业活动中的“销售旺季”和“销售淡季”,农产品和以农产品为原料的某些工业生产的产量和销售量变化,旅游业的“旅游旺季”和“旅游淡季”等。

季节变动有以下三个主要特点。

(1) 季节变动每年重复进行。

(2) 季节变动按照一定的周期进行。需要注意的是,季节变动中的“季节”一词是广义的,它不仅是指一年中的四季,而是泛指任何一种有规律的、按一定周期(季、月、旬、周、日)重复出现的变化。

(3) 每个周期变化强度大体相同。

季节变动的原因通常与自然条件有关,同时也可能由于生产条件、节假日、风俗习惯等社会经济因素所致。季节变动常会给人们的社会经济生活带来某种影响,如会影响某些商品的生产、销售和库存。我们测定季节变动的意义主要在于认识规律、分析过去、预测未来。其目的一是通过分析与预测过去的季节变动规律,为当前的决策提供依据。比如,对一个公司销售活动的研究,可以分析其销售额的变动是季节因素所致,还是由于经营手段或其他偶然因素的影响,从而制定出有效的经营策略。二是对未来现象的季节变动做出预测,以便提前做出合理的安排。三是当需要不包含季节变动因素数据时,能够消除季节变动对序列的影响,以便分析其他构成因素的影响。

二、季节变动的分析方法

季节变动是一种各年变化强度大体相同且每年重现的有规律的变动。测定现象季节变动的主要方法是计算季节比率。季节比率,又称季节指数,是各月(季)平均数与全年总月(季)平均数的比值,它以全期的总平均水平为基准(100%),用百分比形式来反映各月(季)平均水平相对于总平均水平的高低程度。季节比率高说明“旺”,反之说明“淡”。

计算季节指数通常有两种方法:按月(或季)平均法和趋势剔除法。

(一) 按月(或季)平均法

按月(或季)平均法是直接根据原时间序列通过简单平均来计算季节比率的一种常用的方法,该方法的基本思想是,计算出各年同月平均数,以消除随机影响,作为该月(或季)的代表值;然后计算出全部月(或季)的总平均数,作为全年的代表值;再将同月(或同季)平均数与全部月(或季)的总平均数进行对比,即为季节比率。按月(或季)平均法计算季节比率的具体步骤如下。

(1) 根据各年的月份(或季度)数据计算出同月(或同季)的平均数 $\overline{Y}_i$。

(2) 计算出全部数据的总平均数 $\overline{Y}$,找出整个序列的水平趋势。

(3) 计算出各同月(或同季)平均数与总平均数的百分比,即为季节比率(S_i)。其计算公式为

$$S_i = \frac{\overline{Y}_i}{\overline{Y}} \times 100\% \quad (i\text{ 表示月份或者季度}) \tag{6-38}$$

在乘法模型中,季节比率有一个特性,这就是其总和等于季节周期 L(12 或 4),或平均等于 1,即

$$\overline{S} = \frac{\sum S_i}{L} = 1 \tag{6-39}$$

【例 6-14】 某饮料公司 2014—2016 年各月销售量数据如图 6-16 所示。试用按月(或季)平均法计算各月的季节比率。

	A	B	C	D	E	F	G
1–2	月份	各年销售量/万箱			合 计	同月平均 $\overline{Y}_i$	季节比率/%
3		2014	2015	2016	(4)=(1)+(2)+(3)	(5)=(4)÷3	(6)=(5)÷F18
4		(1)	(2)	(3)			
5	1	212.50	255.00	272.00	740	246.50	0.65
6	2	272.00	297.50	306.00	876	291.83	0.77
7	3	314.50	382.50	467.50	1165	388.17	1.02
8	4	331.50	314.50	425.00	1071	357.00	0.94
9	5	357.00	382.50	722.50	1462	487.33	1.28
10	6	467.50	552.50	612.00	1632	544.00	1.43
11	7	357.00	476.00	561.00	1394	464.67	1.22
12	8	340.00	442.00	476.00	1258	419.33	1.10
13	9	306.00	391.00	442.00	1139	379.67	1.00
14	10	280.50	357.00	408.00	1046	348.50	0.92
15	11	263.50	340.00	382.50	986	328.67	0.86
16	12	246.50	323.00	365.50	935	311.67	0.82
17	合计	3749	4514	5440	1 3702	4567	12
18	平均	312.38	376.13	453.33	1 141.83	380.61	1

图 6-16 2014—2016 年各月销售量资料及季节比率计算表

根据上述步骤销售量季节比率计算过程及结果如图 6-16 所示。

为了更清楚地反映饮料销量的季节变动,可以在 Excel 中利用图 6-16 中的季节比率绘制成图形,如图 6-17 所示。可以明显地看出,1—4 月、9—12 月是销售淡季,5—7 月为销售旺季。

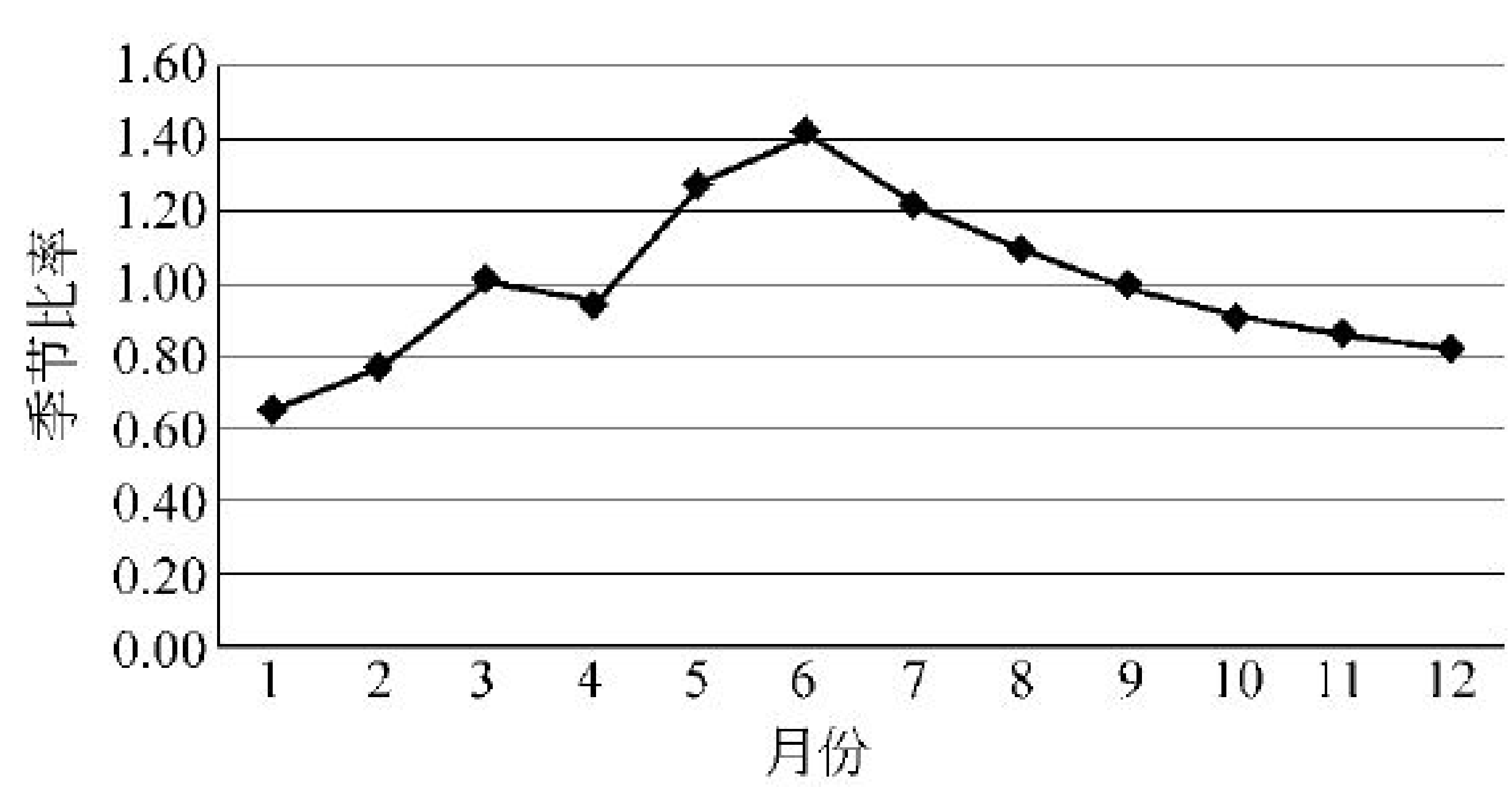


图 6-17 饮料销量的季节变动

按月(或季)平均法计算简单,易于理解。应用该方法的基本假定是:原时间序列没有明显的长期趋势和循环波动,因而,通过若干年同期数值的平均,不仅可以消除不规则波动,而且当平均的周期与循环周期一致时,循环波动也可以在平均过程中得以消除,但实际上,许多时间序列所包含的长期趋势和循环波动,很少能够通过平均予以消除。因此,当时间序列存在明显的长期趋势时,该方法的季节比率不够准确。当存在剧烈的上升趋势时,年末季节指数明显高于年初的季节比率;当存在下降趋势时,年末的季节比率明显低于年初的季节比率。只有当序列的长期趋势和循环波动不明显或影响不重要,可忽略不计时,应用该方法比较合适。

(二) 趋势剔除法

趋势剔除法的基本思想是,先将时间序列中的长期趋势予以消除,然后再计算季节指数。其中,序列中的趋势值可采用移动平均法求得,也可采用最小二乘法求得。利用前者分析季节变动又称为移动平均趋势剔除法,后者简称为趋势剔除法。

采用移动平均趋势剔除法分析季节变动时,假定时间序列各要素的关系结构为:$Y=T\cdot S\cdot C\cdot I$,同时假定各年度的不规则波动 I 彼此独立。由于 12 个月(或 4 个季度)的移动平均数与季节变动的周期(1 年)相同,通过移动平均,可以完全消除季节变动和大部分不规则波动,而仅包含长期趋势和循环波动,结果即为 $T\cdot C$。

然后再将原数列 Y 除以移动平均趋势值 $T\cdot C$,所得百分比称为“季节变动和不规则波动相对数”或“移动平均百分比”,即

$$\frac{T\cdot C\cdot S\cdot I}{T\cdot C}=S\cdot I$$

最后再将各年同月(季)的移动平均百分比加以平均,即可消除不规则波动的影响,只剩下季节变动 S。

具体步骤如下。

(1) 根据各年的月份(或季度)数据,计算 12 个月(或 4 个季度)移动平均趋势值 T。

(2) 将各实际观察值 Y 除以相应趋势值 T,即 $\frac{Y}{T}=S\cdot I$,记为 Y'。

(3) 将 $S\cdot I$ 重新按月(季)排列,求得同月(或同季)平均数 $\overline{Y'_i}$,再将其除以总平均数 $\overline{Y'}$,即得季节比率 S_i。

【例 6-15】 根据图 6-16 的资料,按移动平均趋势剔除法计算销售量的季节比率。

在 Excel 中,首先求出 12 个月移动平均,两两移动平均得到趋势值 T,并求得 Y/T,计算步骤和结果如图 6-18 所示。

然后将图 6-18 中的 Y/T 重新排列,求出各年同月平均数 $\overline{Y'_i}$,使不规则变动消除,进一步计算出季节比率,如图 6-19 所示。由于本题目中 12 个月的季节指数之和等于 1 200%,所以不需要进行调整。

三、季节变动的调整

含有季节变动因素的时间序列,由于受季节影响而产生波动,使序列中的其他特

年/月	销售量Y	12项移动平均	趋势值T	Y'=Y/T
2014年1月	212.50	—	—	—
2	272.00	—	—	—
3	314.50	—	—	—
4	331.50	—	—	—
5	357.00	—	—	—
6	467.50	—	—	—
7	357.00	312.38	314.15	1.14
8	340.00	315.92	316.98	1.07
9	306.00	318.04	320.88	0.95
10	280.50	323.71	323.00	0.87
11	263.50	322.29	323.35	0.81
12	246.50	324.42	327.96	0.75
2015年1月	255.00	331.50	336.46	0.76
2	297.50	341.42	345.67	0.86
3	382.50	349.92	353.46	1.08
4	314.50	357.00	360.19	0.87
5	382.50	363.38	366.56	1.04
6	552.50	369.75	372.94	1.48
7	476.00	376.13	376.83	1.26
8	442.00	377.54	377.90	1.17
9	391.00	378.25	381.79	1.02
10	357.00	385.33	389.94	0.92
11	340.00	394.54	408.71	0.83
12	323.00	422.88	425.35	0.76
2016年1月	272.00	427.83	431.38	0.63
2	306.00	434.92	436.33	0.70
3	467.50	437.75	439.88	1.06
4	425.00	442.00	444.13	0.96
5	722.50	446.25	448.02	1.61
6	612.00	449.79	451.56	1.36
7	561.00	453.33	—	—
8	476.00	—	—	—
9	442.00	—	—	—
10	408.00	—	—	—
11	382.50	—	—	—
12	365.50	—	—	—

图 6-18　销售量季节比率计算(1)

月份	2014年	2015年	2016年	合计	平均	季节比率
1	—	0.76	0.63	1.39	0.69	0.69
2	—	0.86	0.70	1.56	0.78	0.78
3	—	1.08	1.06	2.14	1.07	1.07
4	—	0.87	0.96	1.83	0.92	0.92
5	—	1.04	1.61	2.66	1.33	1.33
6	—	1.48	1.36	2.84	1.42	1.42
7	1.14	1.26	—	2.40	1.20	1.20
8	1.07	1.17	—	2.24	1.12	1.12
9	0.95	1.02	—	1.98	0.99	0.99
10	0.87	0.92	—	1.78	0.89	0.89
11	0.81	0.83	—	1.65	0.82	0.82
12	0.75	0.76	—	1.51	0.76	0.76
合计	—	—	—	23.98	11.99	12.00
平均	—	—	—	2.00	1.00	1.00

图 6-19　销售量季节比率计算(2)

征不能清晰地表现出来,因此,需要将季节变动的影响从时间序列中剔除,以便观察其他特征的影响,这称为季节变动的调整。其方法是将原时间序列除以相应的季节指数,即

$$\frac{Y}{S}=\frac{T\cdot C\cdot S\cdot I}{S}=T\cdot C\cdot I \tag{6-40}$$

结果即为调整后的时间序列,反映了在没有季节因素影响的情况下,时间序列的变化形态。

【例 6-16】 根据图 6-16 的资料,对 2014—2016 年各月的销售量做季节调整,并对调整之后的趋势值用直线方程进行拟合。

在 Excel 中的操作步骤及结果如图 6-20 所示。

	A	B	C	D	E	F
1	年/月	t	销售量Y	季节比率S	季节调整后的序列Y/S	季节调整后的趋势预测值
2	2014年1月	1	213	0.69	305.84	286.36
3	2	2	272	0.78	347.99	291.78
4	3	3	315	1.07	293.00	297.20
5	4	4	332	0.92	361.97	302.63
6	5	5	357	1.33	268.58	308.05
7	6	6	468	1.42	329.32	313.47
8	7	7	357	1.20	297.30	318.90
9	8	8	340	1.12	303.01	324.32
10	9	9	306	0.99	309.18	329.74
11	10	10	281	0.89	314.20	335.16
12	11	11	264	0.82	319.75	340.59
13	12	12	247	0.76	326.00	346.01
14	2015年1月	13	255	0.69	367.01	351.43
15	2	14	298	0.78	380.61	356.86
16	3	15	383	1.07	356.35	362.28
17	4	16	315	0.92	343.41	367.70
18	5	17	383	1.33	287.77	373.12
19	6	18	553	1.42	389.20	378.55
20	7	19	476	1.20	396.40	383.97
21	8	20	442	1.12	393.91	389.39
22	9	21	391	0.99	395.06	394.81
23	10	22	357	0.89	399.90	400.24
24	11	23	340	0.82	412.58	405.66
25	12	24	323	0.76	427.17	411.08
26	2016年1月	25	272	0.69	391.48	416.51
27	2	26	306	0.78	391.48	421.93
28	3	27	468	1.07	435.54	427.35
29	4	28	425	0.92	464.06	432.77
30	5	29	723	1.33	543.56	438.20
31	6	30	612	1.42	431.11	443.62
32	7	31	561	1.20	467.19	449.04
33	8	32	476	1.12	424.21	454.46
34	9	33	442	0.99	446.59	459.89
35	10	34	408	0.89	457.02	465.31
36	11	35	383	0.82	464.15	470.73
37	12	36	366	0.76	483.38	476.16
38		$a=$	280.936 27	$b=$	5.422 772 092	
39	趋势方程		$y=280.936+5.423\times t$			

图 6-20　Excel 中进行季节变动的调整

需要注意的是，图 6-20 中的趋势预测值是不含季节性因素的，也就是说，如果没有季节因素的影响，饮料的销量的预测值在 2016 年 12 月为 476.16 万箱。如果要计算含有季节性的销量的最终预测值，则需要将预测值再乘以该月份的季节比率。

思　考　题

1. 什么是时间序列？它分为哪几种类型？
2. 时间序列构成因素有哪些？
3. 时间序列的编制原则有哪些？
4. 时期序列和时点序列的区别有哪些？
5. 简述定基发展速度与环比发展速度之间的关系。
6. 计算平均发展速度有哪两种方法？它们各具有哪些特点？
7. 简述移动平均法的基本思想。
8. 长期趋势的分析方法有哪些？
9. 如何根据具体的时间序列选择适当的趋势线？
10. 季节变动分析中的按月(季)平均法和趋势剔除法有什么不同？

练　习　题

1. 某企业 2010—2015 年的总产值资料如表 6-8 所示。

表 6-8　某企业 2010—2015 年总产值资料

年　份	2010	2011	2012	2013	2014	2015
总产值/万元	280	320	300	295	360	320

要求：在 Excel 中计算该企业 2010—2015 年的平均总产值。

2. 某企业库存商品的相关统计资料如表 6-9 所示。

表 6-9　某企业库存商品的相关资料

时　间	1 月 1 日	1 月 20 日	2 月 3 日	2 月 25 日	3 月 1 日
库存商品/件	1 800	2 500	1 500	2 800	3 010

要求：在 Excel 中计算该种商品第一季度的平均库存量。

3. 某企业职工人数的统计资料如表 6-10 所示。

表 6-10　某企业职工人数资料

时　间	3 月 31 日	4 月 30 日	5 月 31 日	6 月 30 日
职工人数/人	860	856	860	880

要求：在 Excel 中计算该企业第二季度的平均职工人数。

4. 某工业企业某年第二季度的总产值和职工人数资料如表 6-11 所示。

表 6-11　某工业企业总产值和职工人数资料

月　份	3	4	5	6
总产值/万元	1 500	1 600	1 650	1 850
月末职工人数/人	600	615	630	660

要求在 Excel 中计算：

(1) 第二季度各个月的劳动生产率。

(2) 第二季度月平均劳动生产率。

(3) 第二季度劳动生产率。

5. 根据表 6-7 资料，要求在 Excel 中计算：

(1) 各年逐期增长量、累积增长量及年平均增长量。

(2) 各年环比发展速度、定基发展速度及各自的增长速度。

(3) 按水平法计算年平均发展速度及平均增长速度。

6. 某地区粮食产量 2010—2012 年平均发展速度是 1.03，2013—2015 年平均发展速度是 1.05，2016 年比 2015 年增长 6%，试求 2010—2016 年 7 年的平均发展速度。

7. 某公司过去 18 个月的销售收入数据如表 6-12 所示。

表 6-12 某公司过去 18 个月的销售收入

月份	销售收入/万元	月份	销售收入/万元	月份	销售收入/万元
1	1 033	7	1 334	1	1 572
2	991	8	1 509	2	1 904
3	1 127	9	1 484	3	2 104
4	1 243	10	1 656	4	2 055
5	1 001	11	1 645	5	2 240
6	1 327	12	1 684	6	2 310

要求在 Excel 中操作：用 3 期移动平均法进行趋势分析。

8. 根据表 6-12 资料，要求在 Excel 中操作：

(1) 绘制该时间序列的图形并分析描述其趋势。

(2) 选择一条合适的趋势线拟合数据，写出其趋势方程，计算出各年的趋势值。

9. 某计算机公司 2007—2016 年的计算机销售量数据如表 6-13 所示。

表 6-13 某计算机公司 2007—2016 年的计算机销售量

年份	销售量/千台	年份	销售量/千台	年份	销售量/千台
2007	2	2011	93	2015	106
2008	9	2012	98	2016	108
2009	34	2013	100		
2010	70	2014	105		

要求：根据上述资料分别配合一条指数曲线和 Gompertz 曲线进行趋势分析。

10. 某旅馆的营业收入资料如表 6-14 所示。

表 6-14 某旅馆的营业收入

月份	2014 年	2015 年	2016 年	月份	2014 年	2015 年	2016 年
1	116	145	180	7	2 810	3 120	3 350
2	154	210	245	8	1 204	1 382	1 576
3	220	312	325	9	384	482	625
4	392	520	535	10	183	248	437
5	642	684	710	11	125	130	258
6	1 642	1 872	1 923	12	95	112	166

要求在 Excel 中操作：

(1) 按月平均法计算季节比率。

(2) 按移动平均趋势剔除法计算季节比率，并做出季节变动图。

(3) 对原时间序列做季节性调整，并根据调整后的数据作图。

第七章

指数分析

【统计知识 ABC】

[A]绿色 GDP

人类的经济活动包括两方面的活动。一方面在为社会创造着财富,即所谓“正面效应”;但另一方面又在以种种形式和手段对社会生产力的发展起着阻碍作用,即所谓“负面效应”。这种负面效应集中表现在两个方面,其一是无休止地向生态环境索取资源,使生态资源从绝对量上逐年减少;其二是人类通过各种生产活动向生态环境排泄废弃物或砍伐资源使生态环境从质量上日益恶化。现行的国民经济核算制度只反映了经济活动的正面效应,而没有反映负面效应的影响,因此是不完整的,是有局限性的,是不符合可持续发展战略的。

改革现行的国民经济核算体系,对环境资源进行核算,从现行 GDP 中扣除环境资源成本和对环境资源的保护服务费用,其计算结果可称为“绿色 GDP”。

绿色 GDP 这个指标,实质上代表了国民经济增长的净正效应。绿色 GDP 占 GDP 的比重越高,表明国民经济增长的正面效应越高,负面效应越低,反之亦然。

(资料来源:http://baike.baidu.com/view/41440.htm? fromId=199625)

[B]幸福指数

幸福感是一种心理体验,它既是对生活的客观条件和所处状态的一种事实判断,又是对于生活的主观意义和满足程度的一种价值判断,它表现为在生活满意度基础上产生的一种积极心理体验。而幸福感指数,就是衡量这种感受具体程度的主观指标数值。

如果说 GDP、GNP 是衡量国富、民富的标准,那么,百姓幸福指数就可以成为一个衡量百姓幸福感的标准。

作为社会心理体系一个部分的幸福感,受到许多复杂因素的影响,主要包括:经济因素如就业状况、收入水平等;社会因素如教育程度、婚姻质量等;人口因素如性别、年龄等;文化因素如价值观念、传统习惯等;心理因素如民族性格、自尊程度、生活态度、个性特征、成就动机等;政治因素如民主权利、参与机会等。

2012 年联合国发布了首份《全球幸福指数报告》。根据这份报告,丹麦成为全球最幸福的国家。美国仅排在第 11 名。中国香港第 67 名,中国内地则排名第 112。评价标准极其复杂,财富并非决定性因素,全球总体比以前幸福。

最不幸福国家集中于受贫穷和战火洗礼的非洲国家。较幸福国家倾向较富裕，但收入与幸福并无必然关系；良好精神及身体健康、稳定家庭和婚姻、工作保障等为幸福要素。

（资料来源：http://baike.baidu.com/view/1298996.htm？fromId=497223）

[C]景气指数

将反映各行业运行状况的定量指标，如价格、成交量、开工率等或定性指标，如预期、信心等指数化，来反映经济或行业的景气变化。

常见的景气指数有企业景气指数、经济景气指数。通常景气指数在0～200，100为中间值，高于100视为景气状态，越接近200反映行业经济运行越景气。低于100则是不景气，越接近0则景气越低迷。

（资料来源：http://baike.baidu.com/view/1369057.html.）

企业景气指数（企业综合生产经营景气指数）是根据企业负责人对本企业综合生产经营情况的判断与预期而编制的指数，用以综合反映企业的生产经营状况。

企业景气指数 ＝0.4×即期企业景气指数＋0.6×预期企业景气指数

即期企业景气指数 ＝（企业负责人对本季度本企业综合经营状况回答良好比重－回答不佳的比重）×100＋100

预期企业景气指数 ＝（企业负责人对预计下季度本企业综合经营状况回答良好比重－回答不佳的比重）×100＋100

企业景气指数划分标准为：180以上为“非常景气”区间，[180,150)为“较强景气”区间，[150,120)为“较为景气”区间，[120,110)为“相对景气”区间，[110, 100)为“微景气”区间，100为景气临界点，(100,90]为“微弱不景气”区间，(90,80]为“相对不景气”区间，(80,50]为“较为不景气”区间。

（资料来源：http://baike.baidu.com/view/1369057.html.）

经济景气指数来源于企业景气调查，它是西方市场经济国家建立的一项统计调查制度，是通过对企业家进行定期的问卷调查，并根据企业家对企业经营情况及宏观经济状况的判断和预期来编制的，由此反映企业的生产经营状况、经济运行状况，预测未来经济的发展变化趋势。

经济景气指数目前主要通过两个指标来反映，一是企业家信心指数，是根据企业家对企业外部市场经济环境与宏观政策的认识看法、判断与预期（主要是通过对“乐观”“一般”“不乐观”的选择）而编制的指数，用以综合反映企业家对宏观经济环境的感觉与信心。二是企业景气指数：是根据企业家对本企业综合生产经营情况的判断与预期（主要是通过对“好”“一般”“不佳”的选择）而编制的指数，用以综合反映企业的生产经营状况。景气指数的表示范围为0～200。100为景气指数的临界值，表明景气状况变化不大；100～200为景气区间，表明经济状况趋于上升或改善，越接近200越景气；0～100为不景气区间，表明经济状况趋于下降或恶化，越接近0越不景气。

（资料来源：http://baike.baidu.com/view/1104719.htm.）

第一节　指数的概念及分类

一、指数的概念

指数(index number)是社会经济统计中历史最悠久、应用最广泛、同社会经济生活关系最密切的指标。18 世纪中叶,当时由于美洲新大陆的开发,大量的金银源源不断地流入欧洲,使欧洲的物价飞涨,引起社会不安。当时的经济学家为了测定物价的变动,开始尝试编制物价指数。此后,指数的理论与应用不断发展,应用范围不断扩大,逐渐地运用于社会经济的许多方面。例如,居民消费价格指数、商品零售价格指数等,同人们的日常生活休戚相关;农产品生产价格指数、工业品出厂价格指数、股票价格指数等,则直接影响人们的投资活动。这些指数从不同角度描述了有关经济现象的发展变动程度和轨迹,由于指数在经济分析上具有十分广阔的应用领域,因此,常常被称为"经济指数"。

指数作为一种对比性的统计指标具有相对数的形式,通常表现为百分数,百分数大于或小于 100%,表示变动的升降方向,大于或小于 100%的数额表示变动的升降程度。比如,2016 年我国居民消费价格指数为 102%,表示若以上年的价格水平为 100%,当年全国的价格水平就相当于上年的 102%,或者说价格水平下降了 2.0%。

从对比性质来看,指数通常是不同时间的现象水平的对比,它表明现象在时间上的变动情况(动态)。此外,指数还可以是不同空间(如不同国家、地区、部门、企业等)的现象水平的对比,或者,是现象的实际水平与计划(规划或目标)水平的对比,这些可以看成动态对比指数方法的拓展。

统计指数的概念有广义和狭义两种理解。从广义上讲,就是任何两个数值对比形成的一种相对数。狭义的指数是一种特殊相对数,它是表明由多个个体组成的复杂现象总体数量综合变动程度的相对数。所谓的复杂现象总体,是指由于各个部分的性质不同,在研究数量特征时不能直接相加和直接对比的总体。例如,要说明一个国家或地区商品价格综合变动情况,由于各种商品的经济用途、规格、型号、计量单位等不同,不能直接将各种商品的价格简单对比,而要解决这种复杂现象总体各要素的相加问题,就要编制统计指数综合反映它们的变动情况。本章所讨论的主要是狭义指数的计算原理、编制方法及其在体系分析中的应用。

二、指数的作用

统计指数的作用主要体现在以下几个方面。

(1) 反映复杂现象总体的综合变动方向和变动程度。例如,在统计实践中,经常要研究多种产品的价格的总变动,多种产品产量的总变动等。由于各种产品的使用价值不同,所研究总体中的各个个体不能直接相加。为了分析其总变动,必须利用指数的方法,把不能直接相加总的个体过渡到可以加总对比,从而反映复杂经济现象的总变动方向及变动幅度。

(2) 分析受多因素影响的现象总变动中,各因素变动的影响方向及影响程度。任何

一个复杂现象都是由多个因子构成的，如：销售额＝价格×销售量；原材料费用总额＝产品产量×单耗×原材料单价等，利用指数体系理论可以测定在多种商品销售额的总变动中，价格和销售量的变动对商品销售额变动的影响分别是多少。

（3）分析经济现象总体的长期变化趋势。通过编制指数数列，可以反映现象在较长时间的发展变化趋势。例如，根据2015—2016年的零售商品价格指数，编制成时间数列，可以清楚地观察零售商品价格的变动趋势，研究物价变动对经济建设和人民生活水平的影响程度。

（4）对复杂社会经济现象进行综合评价和测定。随着指数在实际应用中的不断发展和完善，许多经济现象都可以运用统计指数进行综合评价和判断。例如，对综合国力、社会发展水平的综合评价研究。

三、指数的分类

指数的种类很多，根据研究目的不同，可以划分不同的种类。

（一）按所反映现象的特征不同，可分为数量指标指数和质量指标指数

数量指标指数简称数量指数，是反映数量指标变动程度的相对数，如商品销售量指数、产品产量指数等。质量指标指数简称质量指数，是反映质量指标变动程度的相对数，如商品价格指数、产品单位成本指数等。

需要说明的是，数量指标与质量指标的划分具有相对性。如单位产品原材料消耗量指标，相对于产品产量指标，它是质量指标；但相对于单位原材料价格指标，它又是数量指标了。

另外，在统计分析中，我们经常要分析一些总价值量或总实物量指标指数，这些总量指标可以分解为一个数量指标和一个质量指标的乘积，比如销售额等于销售量与价格的乘积，粮食总产量等于收获面积与单位面积产量的乘积。总量指标指数反映了两个指标共同变化的结果，因此在指数分析中，该类指数既不是数量指标指数，也不是质量指标指数。

（二）按所反映的对象范围不同，可分为个体指数和总指数

个体指数是反映总体中个别项目在不同时间上变动的相对数。例如，2016年某地区电视机的产量指数为107.75%，电冰箱产量指数为123.55%等都是个体指数。总指数是反映多个项目组成的总体在不同时间上综合变动程度的相对数。例如，工业总产量指数、零售物价总指数等。

此外，在对总体进行分类或分组的情况下，还需要编制“类（组）指数”，这样在总指数与个体指数之间就产生了类指数。例如，2016年某地区居民消费品价格总指数为99.3%，其中，食品类指数为100.7%、衣着类指数为98%、家庭设备用品类指数为100.2%等，这些都是类指数，类指数实质上也是一种总指数。

(三) 按计算方法或形式不同,可分为简单指数和加权指数

总指数与个体指数之间有一定的联系,可以用个体指数计算相应的总指数。用个体指数简单平均求得的总指数,称为简单指数;用个体指数加权平均求得的总指数,称为加权指数。加权指数是最为常用的计算形式,按照具体计算方法不同又分为加权综合指数和加权平均指数。在本章第二节中将会详细介绍加权指数的编制方法。

(四) 按反映的时间状态不同,可分为动态指数和静态指数

动态指数又叫时间性指数,是由两个不同时期的同类经济变量值对比形成的相对数,反映社会经济现象在不同时间上的发展变化过程和程度。动态指数按所对比的基期的不同,分为定基指数和环比指数两种。常见的居民消费价格指数、股票价格指数等都属于动态指数。

静态指数包括空间指数和计划完成情况指数两种。空间指数(地域指数)指不同空间(如不同国家、地区、部门、企业等)的同类现象水平在同一时间对比的结果,反映现象在不同空间的差异程度。计划完成情况指数则是将同一地区、单位的实际水平与计划水平对比的结果,以反映计划的完成情况。

指数方法论主要论述动态指数,动态指数是出现最早、应用最多的指数,也是理论上最为重要的统计指数。静态指数则是动态指数原理在实际应用中的扩展。

第二节　综合指数和平均指数

一、综合指数

(一) 综合指数的概念和特点

综合指数是将不可同度量的诸经济变量通过另一个有关的同度量因素变量转换成可以相加的总量指标,然后以总量指标对比所得到的相对数。

若所测定的是一组项目的数量变动状况,称为数量指标综合指数,如产品产量指数、商品销售量指数等;若所测定的是一组项目的质量变动状况,则称为质量指标综合指数,如价格指数、产品成本指数等。

综合指数编制时具有以下几个特点。

(1) 借助于同度量因素,先综合后对比。首先将不可直接加总的复杂现象总体的指标,通过引入同度量因素,过渡到能够进行加总的总量指标,然后再进行对比计算相应的综合指数。

例如,表 7-1 中的各种商品的有关数据,在计算价格总指数和销售量总指数时存在两个问题:一是不同商品的价格或销售量是不可以直接加总的,因为直接加总是没有实际意义的。二是用简单综合的方法编制指数明显地受到计量单位的影响。也就是不同商品的价格和销售量都是不可同度量的现象,因而需要找到一个适当的媒介因素,使其转化成可以同度量的形式。如果将每种商品的价格与销售量相乘即可得到销售额,它们是可以

加总的,并且不受计量单位的影响,从分析的角度来看,销售额恰好能反映价格和销售量变动的影响。因此,在编制价格总指数时,可以通过销售量作为媒介因素将价格转化为同度量的销售额形式;同理,在编制多种商品销售量总指数时,可以通过价格作为媒介因素将销售量转换成可以同度量的销售额形式。

表 7-1　某商场三种商品的零售价格和销售量资料

商品名称	计量单位	单价/元		销　售　量	
		2015 年	2016 年	2015 年	2016 年
大米	千克	2.8	3.0	3 100	3 200
猪肉	千克	20	23	800	720
电磁炉	台	360	340	50	65
合计	—	—	—	5 959	5 995

在综合指数的编制过程中,价格和销售量分别起到媒介的作用,将不能直接加总的指标过渡到可以加总,所以称为同度量因素。一般而言,计算数量指数时,应以相应的质量指标作为同度量因素,计算质量指数时,应以相应的数量指标作为同度量因素。

另外,同度量因素还起到对指标加权的作用,因而也被称为综合指数的权数。例如,在计算价格总指数时,同度量因素销售量不仅可以使各种不同商品的销售价格转化为可以加总的价值量,而且由于各种商品的销售量不同,其各自价格变动对综合价格指数影响的大小也有所差别。相应地,具有同度量因素的综合指数就是"加权综合指数"。

(2) 确定同度量因素所属的时期。为了在综合对比过程中单纯反映对比指标的变动或差异程度,即无论是计算数量指数还是质量指数,都要求指数中分子和分母的同度量因素必须是同一时期。既可以都是基期,也可以都是报告期或某一固定时期。使用不同的时期,会产生不同的计算结果,指数的实际意义也会不同。应该固定在哪一时期,通常取决于指数的预期目的和所研究现象的特点。

(3) 综合指数的编制必须以全面资料为基础,且分子、分母所包含的范围必须一致。

(二) 综合指数的编制方法

1. 拉氏指数

拉氏指数是最重要的加权综合指数公式之一,是德国经济统计学家拉斯贝尔斯(Laspeyres)在 1864 年提出的一种指数计算方法,因此又被称为拉氏指数。该公式是将作为权数的同度量因素固定在基期水平上,因此也被称为基期加权综合指数。相应的拉氏质量指数和数量指数的计算公式为

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} \tag{7-1}$$

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} \tag{7-2}$$

式中,I_p 为质量指数;I_q 为数量指数;p_0 和 p_1 分别为一组项目基期和报告期的质量数值;

q_0 和 q_1 分别为一组项目基期和报告期的物量数值。

拉氏质量指数以基期数量指标为权数，目的在于说明假定数量指标保持在基期水平不变的情况下，质量指标的综合变动程度。分子与分母的差额表明因质量指标变动而使总量指标变动(增加或减少)的金额。

拉氏数量指数以基期质量指标为权数，目的在于说明假定质量指标保持在基期水平不变的情况下，数量指标的综合变动程度。分子与分母的差额表明因数量指标变动而使总量指标变动(增加或减少)的金额。

【例 7-1】 某商店 2015 年和 2016 年三种商品的零售价格和销售量资料如表 7-1 所示。试分别以基期销售量和零售价格为权数，计算三种商品的价格综合指数和销售量综合指数。

解：设销售量为 q，零售价格为 p，计算过程如表 7-2 所示。

表 7-2 综合指数计算表

商品名称	计量单位	单价/元		销售量		销售额/元			
		2015 年	2016 年	2015 年	2016 年	2015 年	2016 年	p_0q_1	p_1p_0
		p_0	p_1	q_0	q_1	p_0q_0	p_1q_1		
大米	千克	2.8	3.0	3 100	3 200	8 680	9 600	8 960	9 300
猪肉	千克	20	23	800	720	16 000	16 560	14 400	18 400
电磁炉	台	360	340	50	65	18 000	22 100	23 400	17 000
合计	—	—	—	—	—	42 680	48 260	46 760	44 700

根据式(7-1)，得价格综合指数为

$$I_p = \frac{\sum p_1q_0}{\sum p_0q_0} = \frac{44\ 700}{42\ 680} = 104.73\%$$

根据式(7-2)，得销售量综合指数为

$$I_q = \frac{\sum p_0q_1}{\sum p_0q_0} = \frac{46\ 760}{42\ 680} = 109.56\%$$

计算结果表明，与 2015 年相比，2016 年该商场三种商品的零售价格平均上涨了 4.73%，销售量平均上涨了 9.56%。

可以进一步分析绝对额的变动

$$\sum p_1q_0 - \sum p_0q_0 = 44\ 700 - 42\ 680 = 2\ 020(\text{元})$$

$$\sum p_0q_1 - \sum p_0q_0 = 46\ 760 - 42\ 680 = 4\ 080(\text{元})$$

即由于价格上涨了 4.73%，使销售额增加了 2 020 元；由于销售量增长了 9.56%，使销售额增加了 4 080 元。

拉氏指数由于以基期变量值为权数，可以消除权数变动对指数的影响，从而使不同时期的指数具有可比性。但拉氏指数也存在一定的缺陷。比如，物价指数反映的是在假定基期销售量不变的情况下报告期价格的变动水平，这一指数尽管可以单纯反映价格的变

动水平，但不能反映出消费量的变化。从实际生活角度来看，人们更关心在报告期销售量和销售结构条件下价格变动对实际生活的影响。因此，拉氏价格指数在实际中应用得很少。而拉氏数量指数是假定价格不变的条件下报告期销售量的综合变动，它不仅可以单纯反映出销售量的综合变动水平，也符合计算销售量指数的实际要求。因此，拉氏数量指数在实际中应用得较多。

2. 帕氏指数

帕氏指数也是最重要的加权综合指数公式之一，是德国另一位经济学家帕舍(H. Paasche)在1874年提出的，称为帕氏指数。与拉氏指数不同的是，该公式是将作为权数的同度量因素固定在报告期水平上，故又称为报告期加权综合指数。相应的帕氏质量指数和数量指数的计算公式为

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \tag{7-3}$$

$$I_q = \frac{\sum p_1 q_1}{\sum p_1 q_0} \tag{7-4}$$

式中，I_p 为质量指数；I_q 为数量指数；p_0 和 p_1 分别为一组项目基期和报告期的质量数值；q_0 和 q_1 分别为一组项目基期和报告期的物量数值。

帕氏质量指数以报告期数量指标为权数，目的是说明假定数量指标保持在报告期水平不变的情况下，质量指标的综合变动程度。其分子与分母的差额表明在报告期数量指标条件下，因质量指标的变动而使总量指标变动(增加或减少)的金额。

帕氏数量指数以报告期质量指标为权数，目的是说明质量指标保持在报告期水平不变的情况下，数量指标的综合变动程度。其分子与分母的差额表明在报告期质量指标条件下，由于数量指标的变动而使总量指标变动(增加或减少)的金额。

【例7-2】 根据表7-2中数据资料，分别以报告期销售量和零售价格为权数，计算三种商品的价格综合指数和销售量综合指数。

根据式(7-3)，得价格综合指数为

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{48\,260}{46\,760} = 103.21\%$$

根据式(7-4)，得销售量综合指数为

$$I_p = \frac{\sum p_1 q_1}{\sum p_1 q_0} = \frac{48\,260}{44\,700} = 107.96\%$$

计算结果表明，与2015年相比，2016年该商场三种商品的零售价格平均上涨了3.21%，销售量平均上涨了7.96%。

可以进一步分析绝对额的变动

$$\sum p_1 q_1 - \sum p_0 q_1 = 48\,260 - 46\,760 = 1\,500(\text{元})$$

$$\sum p_1 q_1 - \sum p_1 q_0 = 48\,260 - 44\,700 = 3\,560(\text{元})$$

即由于价格上涨了3.21%，使销售额增加了1 500元；由于销售量增长了7.96%，使

销售额增加了 3 560 元。

帕氏指数将同度量因素固定在报告期,无法消除权数变动对指数的影响,因而不同时期的指数缺乏可比性。在实际应用中,常采用帕氏指数计算质量指数,表明人们在当前的消费结构下需要多支付或少支付的金额,更能反映人们的实际生活。而帕氏数量指数中既有数量指标的变动,也包含了价格的变动(见下式),不符合计算数量指数的目的,因此在实际中应用得较少。

$$\begin{aligned}\sum p_1q_1-\sum p_1q_0&=\sum[(p_1-p_0)+p_0](q_1-q_0)\\&=\sum(p_1-p_0)(q_1-q_0)+\sum p_0(q_1-q_0)\end{aligned}$$

从例 7-1 和例 7-2 的分析中可以看出,采用不同时期的同度量因素作为权数,所得的计算结果是有一定差别的,但从实际应用角度出发,在统计实践中,计算数量指标综合指数时多选用拉氏指数公式(7-2),将作为同度量因素的质量指标固定在基期水平。计算质量指标综合指数时多选用帕氏指数公式(7-3),将作为同度量因素的数量指标固定在报告期水平。

二、平均指数

(一)平均指数的概念和特点

平均指数也是总指数的基本形式之一。综合指数法编制总指数,是将不能直接加总的多个项目,借助同度量因素转换成可以加总的总量指标后再进行对比计算出来的。要求掌握全面的统计资料,但是,有时候却难以取得。而平均指数则弥补了综合指数的这一不足,它是以个体指数为基础,以某一时期的总量为权数对个体指数加权平均计算出来的。

编制平均指数有以下三个特点。

(1) 先对比后平均。首先计算个别现象的个体指数,然后用总量指标作为权数,对个体指数进行加权平均得到总指数。作为权数的总量指标通常是两个变量的乘积,它可以是价值总量,如商品销售额(销售价格与销售量的乘积)、生产费用(单位成本与生产量的乘积),也可以是其他总量,如农产品总产量(单位面积产量与收获面积的乘积)等。

(2) 权数可以是基期或报告期总量指标,也可以采用固定权数。

(3) 可以借助非全面资料计算,也可以采用全面资料。特别是在全面资料不易取得的场合有其特有的使用价值。

(二)平均指数的编制方法

1. 加权算术平均指数

以基期总量为权数对个体指数加权平均计算指数的方法,由于在计算形式上采用了算术平均形式,因此被称为加权算术平均指数。

设基期总量权数为 p_0q_0,个体质量指数为$\frac{p_1}{p_0}$,个体数量指数为$\frac{q_1}{q_0}$,则基期总量加权的

质量指数和数量指数的一般公式为

$$I_p = \frac{\sum \frac{p_1}{p_0} p_0 q_0}{\sum p_0 q_0} \tag{7-5}$$

$$I_q = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0} \tag{7-6}$$

【例 7-3】 利用例 7-1 中的数据资料，如果掌握的资料是非全面数据，如表 7-3 所示，试用基期销售额为权数分别计算价格总指数和销售量总指数。

表 7-3 某商场三种商品的价格和销售量资料

商品名称	计量单位	销售额/元		个体价格指数	个体销售量指数
		2015 年 p_0q_0	2016 年 p_1q_1	p_1/p_0	q_1/q_0
大米	千克	8 680	9 600	1.07	1.03
猪肉	千克	16 000	16 560	1.15	0.90
电磁炉	台	18 000	22 100	0.94	1.30
合计	—	42 680	48 260	—	—

解 根据式(7-5)，得三种商品的价格总指数为

$$I_p = \frac{\sum \frac{p_1}{p_0} p_0 q_0}{\sum p_0 q_0} = \frac{1.07 \times 8\,680 + 1.15 \times 16\,000 + 0.94 \times 18\,000}{42\,680} = 104.52\%$$

根据式(7-6)，得三种商品的销售量总指数为

$$I_q = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0} = \frac{1.03 \times 8\,680 + 0.90 \times 16\,000 + 1.30 \times 18\,000}{42\,680} = 109.51\%$$

计算结果与前面综合指数中的拉氏指数的计算结果完全相同，这是因为当个体指数与总量权数之间存在严格的一一对应关系时，以基期总量为权数计算的加权算术平均指数，实际上是拉氏综合指数的变形。因此，基期总量加权多用于计算数量指数。

2. 加权调和平均指数

以报告期总量为权数对个体指数加权平均计算指数的方法。由于在计算形式上采取了调和平均形式，因此被称为加权调和平均指数。

设报告期总量权数为 p_1q_1，个体质量指数为$\frac{p_1}{p_0}$，个体数量指数为$\frac{q_1}{q_0}$，则报告期总量加权的质量指数和数量指数的一般公式为

$$I_p = \frac{\sum p_1 q_1}{\sum \frac{1}{p_1/p_0} p_1 q_1} \tag{7-7}$$

$$I_q = \frac{\sum p_1 q_1}{\sum \frac{1}{q_1/q_0} p_1 q_1} \tag{7-8}$$

【例 7-4】 根据表 7-3 有关数据，用报告期销售额为权数分别计算价格总指数和销售量总指数。

解 根据式(7-7)，得三种商品的价格总指数为

$$I_p = \frac{\sum p_1 q_1}{\sum \frac{1}{p_1/p_0} p_1 q_1} = \frac{48\,260}{\frac{9\,600}{1.07} + \frac{16\,560}{1.15} + \frac{22\,100}{0.94}} = 102.94\%$$

根据式(7-8)，得三种商品的销售量总指数为

$$I_q = \frac{\sum p_1 q_1}{\sum \frac{1}{q_1/q_0} p_1 q_1} = \frac{48\,260}{\frac{9\,600}{1.03} + \frac{16\,560}{0.90} + \frac{22\,100}{1.30}} = 107.92\%$$

计算结果与前面综合指数中的帕氏指数的计算结果完全相同，这是因为当个体指数与总量权数之间存在严格的一一对应关系时，以报告期总量为权数计算的加权调和平均指数，实际上是帕氏综合指数的变形。因此，报告期总量加权多用于计算质量指数。

通过例 7-4 可以看出，平均指数实际上是综合指数的一种变形。但二者所依据的计算资料是不同的。综合指数的计算通常需要掌握全面的资料，实际编制中往往具有一定的困难，而平均指数则既可以依据全面的资料来编制，也可以依据非全面资料来编制，因此平均指数在实际中应用得更为广泛。此外，平均指数中的权数也可以采取比重形式，其权数(W)可以在一定时期内相对固定下来，连续使用几年，这就是所谓的固定权数加权的平均指数。例如，我国的商品零售价格指数和居民消费价格指数就是采用固定权数加权的算术平均形式计算的，其权数每年根据住户调查资料做相应的调整。

第三节　指数体系与因素分析

一、指数体系的概念及作用

社会经济现象之间的相互联系、相互影响的关系是客观存在的。这种经济表现为指标间的数量对等关系，如：

商品销售额 = 商品销售量 × 商品销售价格

生产总成本 = 产品产量 × 单位产品成本

原材料费用总额 = 产品产量 × 单位产品原材料消耗量 × 单位原材料价格

销售利润 = 销售量 × 销售价格 × 销售利润率

上述的这种关系，按指数形式表现时，同样也存在这种对等关系。即

商品销售额指数=商品销售量指数×商品销售价格指数

生产总成本指数=产品产量指数×单位产品成本指数

原材料费用总额指数=产品产量指数×单耗指数×单位原材料价格指数

销售利润指数=销售量指数×销售价格指数×销售利润率指数

在统计分析中,通常把由三个或三个以上相互联系的指数构成的体系称为指数体系。指数体系在统计分析中的作用有以下三点。

(1) 对复杂经济现象进行因素分析。利用指数体系可以分析在复杂经济现象总变动中,各因素的变动对总变动的影响方向和程度。

(2) 进行指数推算。即利用各指数之间的联系,根据已知的指数推算未知的指数。例如,根据已知的商品销售额指数和价格指数,可以推算出销售量指数。即销售量指数=销售额指数÷价格指数。

(3) 利用指数体系既可进行总量指标变动的因素分析,也可进行平均指标变动的因素分析。比如,分析某企业的劳动生产率指数时,可以根据其影响因素,分解分析各车间职工人数变动与各车间劳动生产率变动对总指数的影响方向和程度。

二、总量指标变动的因素分析

总量指标可以表现为价值总量(如总产值、总成本、销售额等),也可以表现为实物总量(如农作物总产量、原材料消耗总量等)。不同时期总量指标对比形成的相对数称为总量指标指数(简称总量指数)。由总量指数及其若干个因素指数构成的数量关系式称为总量指数体系。总量指标指数体系一般具有以下几个特征。

(1) 在各因素指数中,一个为数量指数,另一个必须为质量指数。

(2) 总量指数等于各因素指数的乘积。

(3) 总量的变动差额等于各个因素变动差额之和。

(4) 各因素指数的权数必须是不同时期的,即数量指数的权数应固定在基期,质量指数的权数应固定在报告期。

对总量指标进行因素分析时,首先要将其进行因素分解,根据因素的多少不同,可以采用两因素分析或多因素分析。

(一) 总量指标变动的两因素分析

如果总量指标可以分解为一个数量指标和一个质量指标的乘积,则总量指数可分解为数量指标综合指数和质量指标综合指数两因素的乘积。指数体系如下:

$$\frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \cdot \frac{\sum p_1 q_1}{\sum p_0 q_1} \tag{7-9}$$

绝对额关系如下:

$$\sum q_1 p_1 - \sum q_0 p_0 = \left(\sum q_1 p_0 - \sum q_0 p_0\right) + \left(\sum p_1 q_1 - \sum p_0 q_1\right) \tag{7-10}$$

【例 7-5】 某工业企业生产三种产品,产量及单位成本等有关数据如表 7-4 所示,利用指数体系分析单位成本和产量变动对总成本的影响。

根据表 7-4 中资料计算:

总成本指数:$\dfrac{\sum q_1 p_1}{\sum q_0 p_0} = \dfrac{270}{236} = 114.41\%$

表 7-4 某工业企业三种产品的有关数据

产品名称	计量单位	产品产量		单位成本/元		总成本/万元		
		q_0	q_1	p_0	p_1	q_0p_0	q_1p_1	q_1p_0
A	箱	5 000	5 500	200	204	100	112.2	110
B	台	12 000	14 000	50	60	60	84	70
C	件	38 000	41 000	20	18	76	73.8	82
合计	—	—	—	—	—	236	270	262

总成本增加的绝对额：$\sum q_1p_1 - \sum q_0p_0 = 270 - 236 = 34$(万元)

其中：

产品产量变动的影响：$\dfrac{\sum q_1p_0}{\sum q_0p_0} = \dfrac{262}{236} = 111.02\%$

产品产量增加使总成本增加的绝对额：$\sum q_1p_0 - \sum q_0p_0 = 262 - 236 = 26$(万元)

产品单位成本变动的影响：$\dfrac{\sum p_1q_1}{\sum p_0q_1} = \dfrac{270}{262} = 103.05\%$

产品单位成本提高使总成本增加的绝对额：

$$\sum p_1q_1 - \sum p_0q_1 = 270 - 262 = 8(\text{万元})$$

用相对数表示：114.41%＝111.02%×103.05%

用绝对额表示：34 万元＝26 万元＋8 万元

综上所述，该工业企业报告期三种产品总成本比基期增长了 14.41%，增加额为 34 万元，是由于产品产量和产品单位成本两因素发生变动共同引起的，其中产品产量增长了 11.02%，使总成本增加了 26 万元，产品单位成本增长了 3.05%，使总成本增加了 8 万元。

(二) 总量指标的多因素分析

如果总量指数可以分解为三个或三个以上的因素，需要进行多因素变动分析。多因素现象的指标体系，由于所包含的因素较多，指数的编制过程比较复杂，因此应注意以下几个问题。

1. 各因素的排列顺序

在多因素分析中，各因素的排列应有一定顺序，因素现象中的指标有数量指标和质量指标，数量指标是基础，是前提，质量指标是数量指标的派生指标。所以，应按照先数量指标后质量指标的顺序排列。

2. 相邻因素相乘应有明确的经济意义

各因素的具体排列顺序除了以上要求外，还要考虑每个因素之间的内在联系，使相邻的两个因素相乘都具有明确的经济意义。例如，在例 7-6 中，研究工业生产原材料费用总额变动时，各影响因素的排列顺序为产品产量、单位产品原材料消耗量和单位原材料价格，即

原材料费用总额＝产品产量×单耗×单位原材料价格

上述公式中，产品产量与单耗的乘积为原材料消耗总量，具有经济意义；单耗与单位原材料价格的乘积为单位原材料的消耗额，也具有经济意义。可见排列顺序是合理的。

3. 确定同度量因素的时期

在编制多因素指标所组成的综合指数时，为测定某一因素指标的变动影响，要把其中两个或两个以上的因素固定不变。采取逐项固定的方法，即将未被分析的后面诸因素固定在基期水平，而已被分析过的前面诸因素，则固定在报告期水平。其实原则上仍是遵循综合指数编制的一般要求，即编制数量指标指数时，以基期的质量指标为同度量因素；编制质量指标指数时，以报告期的数量指标为同度量因素。仍以例 7-6 为例，分析产品产量指数时，单耗与单位原材料价格都是质量指标，应固定在基期；分析单耗指数时，产品产量是数量指标，应固定在报告期，单位原材料价格是质量指标，应固定在基期；在分析单位原材料价格指数时，产品产量×单位产品原材料消耗量＝原材料消耗总量，变成数量指标，应固定在报告期。

以工业生产原材料费用总额的变动为例。

原材料费用总额＝产品产量×单位产品原材料消耗量×单位原材料价格

由此建立的多因素指数体系为

原材料费用总额指数＝产品产量指数×单耗指数×原材料单价格指数

用符号表示为

$$\frac{\sum q_1 m_1 p_1}{\sum q_0 m_0 p_0} = \frac{\sum q_1 m_0 p_0}{\sum q_0 m_0 p_0} \cdot \frac{\sum q_1 m_1 p_0}{\sum q_1 m_0 p_0} \cdot \frac{\sum q_1 m_1 p_1}{\sum q_1 m_1 p_0} \tag{7-11}$$

绝对额为

$$\sum q_1 m_1 p_1 - \sum q_0 m_0 p_0 = \left(\sum q_1 m_0 p_0 - \sum q_0 m_0 p_0\right) + \left(\sum q_1 m_1 p_0 - \sum q_1 m_0 p_0\right) + \left(\sum q_1 m_1 p_1 - \sum q_1 m_1 p_0\right) \tag{7-12}$$

式中，q 为产品产量；m 为单位产品原材料消耗量；p 为单位原材料价格；0，1 分别为基期和报告期。

【例 7-6】 设有某企业三种产品的产量、单耗和原材料单价的相关资料，分析原材料费用总额的变动受产品产量、单耗和原材料单价变动的影响分别是多少。

解 根据表 7-5，可以计算三种产品的原材料费用总额，以进一步分析原材料费用总额的变动情况及其原因，如表 7-6 所示。

表 7-5　三种产品的产量、单耗及原材料单价情况

产品名称	产品产量		单位产品原材料消耗量/千克		单位原材料价格/元	
	基期	报告期	基期	报告期	基期	报告期
	q_0	q_1	m_0	m_1	p_0	p_1
A	200	230	8	9	100	106
B	300	280	10	10	30	32
C	500	550	30	28	40	43

表 7-6　三种产品原材料费用总额计算表

产品名称	原材料费用总额/万元			
	$q_1m_1p_1$	$q_0m_0p_0$	$q_1m_0p_0$	$q_1m_1p_0$
A	21.94	16.00	18.40	20.70
B	8.96	9.00	8.40	8.40
C	66.22	60.00	66.00	61.60
合计	97.12	85.00	92.80	90.70

原材料费用总额的变动情况，即

$$原材料费用总额指数 = \frac{\sum q_1m_1p_1}{\sum q_0m_0p_0} = \frac{97.12}{85} = 114.26\%$$

原材料费用变动额：$\sum q_1m_1p_1 - \sum q_0m_0p_0 = 97.12 - 85 = 12.12$(万元)

(1) 产品产量变动的影响

$$产品产量指数 = \frac{\sum q_1m_0p_0}{\sum q_0m_0p_0} = \frac{92.8}{85} = 109.18\%$$

产品产量变动对总费用变动影响的绝对额：

$$\sum q_1m_0p_0 - \sum q_0m_0p_0 = 92.8 - 85 = 7.8(万元)$$

(2) 单位产品原材料消耗量变动的影响

$$产品单耗指数 = \frac{\sum q_1m_1p_0}{\sum q_1m_0p_0} = \frac{90.7}{92.8} = 97.74\%$$

产品单耗变动对总费用变动影响的绝对额：

$$\sum q_1m_1p_0 - \sum q_1m_0p_0 = 90.7 - 92.8 = -2.1(万元)$$

(3) 单位原材料价格变动的影响

$$单位原材料价格指数 = \frac{\sum q_1m_1p_1}{\sum q_1m_1p_0} = \frac{97.12}{90.7} = 107.08\%$$

单位原材料价格变动对总费用变动影响的绝对额：

$$\sum q_1m_1p_1 - \sum q_1m_1p_0 = 97.12 - 90.7 = 6.42(万元)$$

相对数的关系为：114.26%＝109.17%×97.74%×107.08%

绝对额的关系为：12.12＝7.8＋(－2.1)＋6.42

综上所述，原材料费用总额由基期的 85 万元上升至报告期的 97.12 万元，增长了 14.26%，增加了 12.12 万元。这一结果是由于产量、单耗、单位原材料价格三个因素共同影响的结果。其中，产量增长 9.17%，使费用总额增加 7.8 万元；单耗降低 2.26%，使费用总额减少 2.1 万元；单位原材料价格增长 7.08%，使费用总额增加 6.42 万元。

三、平均指标指数的因素分析

平均指标指数是两个不同时期的平均指标对比计算的相对数，用以反映现象平均水

平的变动。平均指标通常是在分组数据条件下，用加权算术平均方式计算的，通过对加权算术平均数的分解，可以分析影响平均数变动的各因素。如下式：

$$\bar{x}=\frac{\sum xf}{\sum f}=\sum x\cdot\frac{f}{\sum f}$$

上式说明，平均指标的变动实际上受两个因素的影响：一个是各组平均指标变动的影响，另一个是各组单位数在总体中所占比重变动的影响。因此，我们可以运用因素分析法来分析这两个因素变动对平均指标变动的影响方向和影响程度。

根据总量指标指数因素分析方法的要求，对于平均指标的两因素分析，首先要构建一个平均指标指数体系，可按照如下方式定义有关平均指标指数。

（一）平均指标指数

平均指标指数也称为可变构成指数，它反映了平均指标的实际变动情况。其计算公式为

$$\text{平均指标指数}=\frac{\bar{x}_1}{\bar{x}_0}=\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}\tag{7-13}$$

式中，$\bar{x}$ 为总平均指标；x 为各组标志值，即平均水平；f 为各组单位数。

（二）固定构成指数

固定构成指数指将总体中各组单位数所占比重即总体结构固定起来计算的平均指标指数，单纯反映各组平均水平的变动对总平均指标变动的影响。其计算公式为

$$\text{固定构成指数}=\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_1}{\sum f_1}\tag{7-14}$$

（三）结构影响指数

结构影响指数指将各组平均水平固定起来计算的平均指标指数，单纯反映总体结构变动对总平均指标变动的影响。其计算公式为

$$\text{结构影响指数}=\frac{\sum x_0f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}\tag{7-15}$$

由上述方法定义的有关平均指标指数，构成如下平均指标指数体系，从相对量角度：

$$\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}=\left(\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_1}{\sum f_1}\right)\cdot\left(\frac{\sum x_0f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}\right)\tag{7-16}$$

即平均指标指数＝固定构成指数×结构影响指数。

从绝对量角度：

$$\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}=\left(\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_1}{\sum f_1}\right)+\left(\frac{\sum x_0f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}\right)\tag{7-17}$$

即平均指标的增加额＝由于变量水平的变动引起的平均指标的增加额＋由于结构的变动引起的平均指标的增加额。

上述公式是对平均指标的变动进行因素分析的基础。

【例 7-7】 某公司职工人数和月平均工资的分组资料如表 7-7 所示。试对该公司职工平均工资的变动进行因素分析。

表 7-7 某公司员工工资情况表

工人类别	月均工资/元		职工人数/人		工资总额/元		
	基期(x_0)	报告期(x_1)	基期(f_0)	报告期(f_1)	(x_0f_0)	(x_1f_1)	(x_0f_1)
一车间	900	930	220	270	198 000	251 100	243 000
二车间	1 120	1 150	115	130	128 800	149 500	145 600
三车间	1 460	1 510	135	100	197 100	151 000	146 000
合　计	—	—	470	500	523 900	551 600	534 600

(1) 计算平均工资指数,说明该公司平均工资的变动情况。

$$\text{报告期的平均工资：}\bar{x}_1=\frac{\sum x_1f_1}{\sum f_1}=\frac{551\,600}{500}=1\,103.2(\text{元})$$

$$\text{基期的平均工资：}\bar{x}_0=\frac{\sum x_0f_0}{\sum f_0}=\frac{523\,900}{470}=1\,114.68(\text{元})$$

$$\text{平均工资指数}=\frac{\bar{x}_1}{\bar{x}_0}=\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}=\frac{1\,103.2}{1\,114.68}=98.97\%$$

变动的绝对额：$\bar{x}_1-\bar{x}_0=1\,103.2-1\,114.68=-11.48$(元)

(2) 计算固定构成指数,说明各组工资水平变动对总平均工资的影响。

$$\text{固定构成指数}=\frac{\sum x_1f_1}{\sum f_1}\div\frac{\sum x_0f_1}{\sum f_1}=\frac{551\,600/500}{534\,600/500}=\frac{1\,103.2}{1\,069.2}=103.18\%$$

$$\text{变动的绝对额：}\frac{\sum x_1f_1}{\sum f_1}-\frac{\sum x_0f_1}{\sum f_1}=1\,103.2-1\,069.2=34(\text{元})$$

(3) 计算结构影响指数,说明各组构成变化对总平均工资变动的影响。

$$\text{结构影响指数}=\frac{\sum x_0f_1}{\sum f_1}\div\frac{\sum x_0f_0}{\sum f_0}=\frac{534\,600/500}{523\,900/470}=\frac{1\,069.2}{1\,114.68}=95.92\%$$

$$\text{变动的绝对额：}\frac{\sum x_0f_1}{\sum f_1}-\frac{\sum x_0f_0}{\sum f_0}=1\,069.2-1\,114.68=-45.48(\text{元})$$

各指数之间的相对数关系为：98.97%＝103.18%×95.92%

各因素影响差额之间的关系为：－11.48＝34＋(－45.48)

上述结果表明：报告期与基期相比,虽然三个车间工人的月平均工资都增加了,但是

总平均工资却下降了1.03%，减少了11.48元。其中，由于各车间平均工资提高，使总平均工资提高了3.18%，增加了34元；但由于各车间工人人数的结构发生了较大变化，收入水平最高的三车间的人数占全部工人数的比重由基期的28.72%，下降到报告期的20%，从而使总平均工资下降了4.08%，减少了45.48元。

四、含总平均指标变动的总量指标变动的因素分析

（一）含总平均指标变动的总量指标变动的两因素分析

平均指标的变动往往是总量指标变动的一个重要因素。因此，在分析总平均指标变动的绝对额时，还可进一步分析由于总平均指标的变动引起的总量指标变动的绝对变动情况。将两个时期的平均指标的差额乘以报告期的总体单位数（$\sum f_1$）即可确定。

例如，例7-7中由于总平均工资发生变动对工资总额的影响为

$$(\bar{x}_1-\bar{x}_0)\cdot\sum f_1=(1\,103.2-1\,114.68)\times 500=-5\,740(\text{元})$$

结果表明，由于总平均工资报告期比基期减少了11.48元，使报告期工资总额减少5 740元。其中：

(1) 由于各组工人月均工资的变动使工资总额增加：

$$\left(\frac{\sum x_1 f_1}{\sum f_1}-\frac{\sum x_0 f_1}{\sum f_1}\right)\cdot\sum f_1=(1\,103.2-1\,069.2)\times 500=17\,000(\text{元})$$

(2) 由于各组职工人数结构的变动使工资总额减少：

$$\left(\frac{\sum x_0 f_1}{\sum f_1}-\frac{\sum x_0 f_0}{\sum f_0}\right)\cdot\sum f_1=(1\,069.2-1\,114.68)\times 500=-22\,740(\text{元})$$

（二）含总平均指标变动的总量指标变动的多因素分析

含总平均指标变动的总量指标变动的多因素分析是将总平均指标指数和总量指标指数结合起来进行分析，进一步分析总量指标变动中受总平均指标变动的影响情况。

根据总平均指标和总量指标间存在的经济联系，得出总平均指标指数体系与总量指标指数体系间必然的联系，然后，再进行分析。以工资总额指数体系为例。

工资总额指数＝职工人数指数×总平均工资指数

用公式表示为

$$\frac{\sum x_1 f_1}{\sum x_0 f_0}=\frac{\sum f_1}{\sum f_0}\cdot\left(\frac{\sum x_1 f_1}{\sum f_1}\div\frac{\sum x_0 f_0}{\sum f_0}\right)\tag{7-18}$$

由于总平均工资指数＝固定构成指数×结构影响指数

则

工资总额指数＝职工人数指数×固定构成指数×结构影响指数

用公式表示为

$$\frac{\sum x_1 f_1}{\sum x_0 f_0}=\frac{\sum f_1}{\sum f_0}\cdot\left[\left(\frac{\sum x_1 f_1}{\sum f_1}\div\frac{\sum x_0 f_1}{\sum f_1}\right)\cdot\left(\frac{\sum x_0 f_1}{\sum f_1}\div\frac{\sum x_0 f_0}{\sum f_0}\right)\right]\tag{7-19}$$

绝对数关系式为

$$\sum x_1 f_1 - \sum x_0 f_0 = \left(\sum f_1 - \sum f_0\right)\cdot \bar{x}_0 + \left(\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_1}{\sum f_1}\right)\cdot \sum f_1 + \left(\frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0}\right)\cdot \sum f_1 \tag{7-20}$$

【例 7-8】 以表 7-7 数据为例，计算该公司工资总额指数，并分析工资总额受各因素影响的情况。

解 利用式(7-19)和式(7-20)，分别从相对数和绝对数两方面进行分析。

(1) 相对数分析。

$$\text{工资总额指数} = \frac{\sum x_1 f_1}{\sum x_0 f_0} = \frac{551\ 600}{523\ 900} = 105.29\%$$

其中：

$$① \text{职工人数指数} = \frac{\sum f_1}{\sum f_0} = \frac{500}{470} = 106.38\%$$

$$② \text{固定构成指数} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} = \frac{1\ 103.2}{1\ 069.2} = 103.18\%$$

$$③ \text{结构影响指数} = \frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} = \frac{1\ 069.2}{1\ 114.68} = 95.92\%$$

相对数关系式：105.29%＝106.38%×103.18%×95.92%

(2) 绝对数分析。

工资总额变动额：551 600－523 900＝27 700(元)

其中：

① 职工人数增加使工资总额增加：

$$\left(\sum f_1 - \sum f_0\right)\cdot \bar{x}_0 = (500 - 470) \times \frac{523\ 900}{470} = 33\ 440.43(\text{元})$$

② 由于各车间平均工资增加使工资总额增加：

$$\left(\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_1}{\sum f_1}\right)\cdot \sum f_1 = (1\ 103.2 - 1\ 069.2) \times 500 = 17\ 000(\text{元})$$

③ 由于职工人数结构变动使工资总额减少：

$$\left(\frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0}\right)\cdot \sum f_1 = \left(\frac{524\ 600}{500} - \frac{523\ 900}{470}\right) \times 500 = -22\ 740.43(\text{元})$$

绝对数关系式：27 700＝33 440.43＋17 000＋(－22 740.43)

计算结果表明：该公司工资总额增长了 5.29%，增加了 27 700 元，是由于职工人数、各车间平均工资水平和职工人数构成三因素共同影响的结果。其中由于职工人数的增加使其增长了 6.38%，增加了 33 440.43 元；由于三个车间的平均工资变动使其增长了 3.18%，增

加了 17 000 元；但由于各车间人数构成变动使其下降了 4.08％，减少了 22 740.43 元。

第四节　几种常用的经济指数

指数作为一种重要的经济分析指标和方法，在实践中得到了广泛应用。从我国现行的统计调查制度来看，目前编制的价格指数主要有：居民消费价格指数（CPI），商品零售价格指数（retail price index，RPI），农业生产资料价格指数（agriculture manufacture price index，AMPI），生产者价格指数（producer price index，PPI），原材料、燃料、动力购进价格指数，固定资产投资价格指数，房地产价格指数，农产品收购价格指数，等等。此外，还有部分行业结合实际，编制了其他价格指数。例如：海关编制了进出口价格指数。现以目前国内外常见的主要经济指数为例，对指数方法的具体应用加以介绍。

一、居民消费价格指数

居民消费价格指数在国外称之为消费者价格指数，是反映一定时期内城乡居民所购买的生活消费品价格和服务项目价格变动趋势和程度的相对数，是对城市居民消费价格指数和农村居民消费价格指数进行综合汇总计算的结果。利用居民消费价格指数，可以反映通货膨胀或通货紧缩程度，借以观察和分析消费品的零售价格和服务价格变动对城乡居民实际生活费支出的影响程度，为研究居民生活、进行宏观经济分析与决策和进行价格总水平的监测和调控提供依据。

编制居民消费价格指数的资料采用抽样调查和重点调查相结合的方法取得，即在全国选择不同经济区域和分布合理的地区，以及有代表性的商品作为样本，对其市场价格进行定期调查，以样本推断总体。编制过程按下列几个步骤进行。

(1) 选择调查地区和调查点。按照大中小兼顾以及地区分布合理原则，采用划类选择法抽选调查市县约 500 个。依据经济普查获得的企业名录库，以零售额或经营规模为标志，从高到低排队随机等距抽样，兼顾规模、分布、注册登记类型等因素适当调整，抽选 6.3 万个调查网点（包括食杂店、百货店、超市、便利店、专业市场、专卖店、购物中心、农贸市场、服务消费单位等）。

(2) 选择代表商品和代表规格品。代表商品是选择那些消费量大、质量好、货源稳、价格变动有代表性的商品；代表规格品的确定是根据商品零售资料和 5.9 万户城市居民、6.8 万户农村居民的消费支出记账资料，按照有关规定筛选的。

目前，居民消费价格调查按用途划分为食品、烟酒及用品、衣着、家庭设备用品及维修服务、医疗保健和个人用品、交通和通信、娱乐教育文化用品和服务、居住八大类商品，从以上分类中又选定 262 个代表性商品的基本分类和 600～700 种代表规格品入编指数。

(3) 价格调查方式。全国有 4 000 余名现场调查员，按照定人、定点、定时原则，直接对采价网点的代表规格品进行价格采集。同时全国聘请近万名辅助调查员协助登记调查。对价格变动不频繁的商品和服务项目，如汽车、邮资、物管、报刊、医疗等每月采价 1 次；对于烟酒、家电、服装等一般性调查商品和服务项目每月采价 2～3 次；对于居民生

活密切相关、价格变动比较频繁的主副食品价格,如蔬菜、肉、禽、水果,每 5 天采集 1 次价格。

(4) 权数的确定。权数的计算依据是全国 13.3 万户城乡居民家庭收支抽样调查资料,并辅之以典型调查数据做补充。一般每 5 年更换一次,5 年期内各年度适当调整。

(5) 编制价格指数。我国现行居民消费价格指数主要是采用固定加权算术平均指数方法编制的。其计算公式为

$$I_p = \frac{\sum I_i \cdot w}{\sum w} \tag{7-21}$$

式中,I_i 为类指数;w 为固定权数。

在编制各层次的消费价格指数时,采用的公式有所不同。

① 小类指数,即代表规格品的环比价格指数:$i=\frac{p_1}{p_0}$。

② 基本分类指数,是各代表规格品环比价格指数的几何平均数:$i_p=\sqrt[n]{i_1 \cdot i_2 \cdot i_3 \cdot \cdots \cdot i_n}$。

③ 中类指数,以基本分类指数乘以相应的权数计算得到,计算公式为 $I_p=\frac{\sum i_p \cdot w}{\sum w}$。

④ 大类指数和总指数,以各中类指数乘以相应的权数,便得到大类指数。大类指数乘以相应的权数,便得到总指数。计算公式同中类指数。

【例 7-9】 给出居民消费价格指数计算表(表 7-8)。已知各大类、交通和通信中类及其代表规格品的有关资料,要求据以编制有关的价格指数。

表 7-8 居民消费价格指数计算表

类别及品名	规格等级	计量单位	平均价格/元		指数/%	权数	指数×权数
			基期	计算期			
总指数	—	—	—	—	104.89	100	—
一、食品	—	—	—	—	108.15	42	45.423
二、烟酒及用品					108.74	6	6.524
三、衣着	—	—	—	—	98.46	15	14.769
四、家庭设备用品及服务	—	—	—	—	102.70	11	11.297
五、医疗保健及个人用品	—	—	—	—	110.43	3	3.313
六、交通和通信	—	—	—	—	100.34	4	4.013
1. 交通工具(中类)	—	—	—	—	103.45	[60]	62.068
摩托车(基本分类)		—	—	—	101.54	＜45＞	45.693
自行车(基本分类)		—	—	—	105.29	＜50＞	52.643
	凤凰	辆	360	386	107.22	—	—
	飞鸽	辆	532	550	103.38	—	—
三轮车(基本分类)		—			102.22	＜5＞	5.111
2. 通信工具(中类)	—	—	—	—	95.67	[40]	38.268
七、娱乐教育文化用品及服务	—	—	—	—	101.26	5	5.063
八、居住项目	—	—	—	—	103.50	14	14.490

解　利用表中资料和公式，依次计算个体指数、各类别指数和消费价格总指数如下。

① 计算小类指数。

$$凤凰自行车价格指数=\frac{386}{360}=107.22\%$$

$$飞鸽自行车价格指数=\frac{550}{532}=103.38\%$$

② 计算基本分类指数，如自行车指数。

$$i_p=\sqrt[2]{107.22\%\times103.38\%}=105.29\%$$

③ 计算中类指数，如交通工具指数。

$$I_p=\frac{\sum i_p\cdot w}{\sum w}=\frac{45.693+52.643+5.111}{100}=103.45\%$$

④ 计算大类指数，如交通和通信指数。

$$I_p=\frac{\sum i_p\cdot w}{100}=\frac{62.068+38.268}{100}=100.34\%$$

⑤ 计算居民消费价格总指数。

$$\begin{aligned}I_p&=\frac{\sum i_p\cdot w}{\sum w}\\&=\frac{45.423+6.524+14.769+11.297+3.313+4.013+5.063+14.490}{100}\\&=104.89\%\end{aligned}$$

二、商品零售价格指数

商品零售价格指数是反映一定时期内城乡商品零售价格变动趋势和程度的相对数。零售物价的调整变动直接影响到城乡居民的生活支出和国家的财政收入，影响居民购买力和市场供需平衡，影响消费与积累的比例。因此，编制商品零售价格指数可以掌握市场商品价格的变动状况，并在此基础上编制其他各种派生指数，为研究市场流通和进行国民经济核算提供依据。

我国的零售物价指数编制程序与消费者价格指数基本相同：首先，选择部分地区作为代表区域，就所选定的代表品收集价格数据，采集价格资料时，应区分不同市场选择若干个价格调查点。目前在全国共抽选出226个市县作为国家编制零售物价指数的价格调查点，其中城市146个、县80个。其次，将零售商品按用途划分为16个大类、229个基本分类、500种以上的规格产品。其中不包括服务项目。再次，在所采集价格资料的基础上，先计算该品种的平均价格，而后按“市县→省→全国”的顺序分级编制计算各单项商品的零售价格指数和类别指数。编制商品零售价格指数所用的权数：大类权数根据批发零售贸易统计中的相关资料和其他相关资料推算，小类及基本分类的权数参考居民消费价格指数中的相关权数和有关典型调查资料确定。最后，采用固定加权算术平均指数公式编制零售商品价格总指数。

三、生产者价格指数

生产者价格指数是从生产者角度考虑的物价指数,反映生产者在初级市场(非零售市场)上出售或购买的产品的价格变动程度。广义的生产者价格指数包括有关国民经济各产业的原材料、半成品和产成品三个环节的价格指数。狭义的生产者价格指数仅指工农业等的产品价格指数。在我国,生产者价格指数通常是指工业品出厂价格指数,也可包括农产品生产价格指数。

我国的工业品出厂价格指数是反映一定时期内全部工业产品出厂价格总水平的变动趋势和程度的相对数,包括工业企业售给本企业以外所有单位的各种产品和直接售给居民用于生活消费的产品。农产品生产价格指数是反映一定时期内,农产品生产者出售农产品价格水平变动趋势及幅度的相对数。这两种指数可以客观反映相关产业的产品生产价格水平和结构变动情况,满足工农业统计核算乃至整个国民经济核算和宏观经济分析的需要。

这些价格指数的编制方法与前面所述的“居民消费价格指数”和“商品零售价格指数”大致相似,区别在于涉及的产品范围、产品分类方式以及产品计价标准等有所不同。例如,目前,我国工业品出厂价格指数调查的产品有 4 000 多种(含规格品 9 500 多种),覆盖全部 39 个工业行业大类,涉及调查种类 186 个。编制工业品出厂价格指数所用的权数用工业品销售额计算,计算资料来源于工业普查数据。权数一般五年更换一次,与国外通行的做法一样,在五年期间,若出现产品更新换代快,以致影响权数代表性的情况时,可及时进行合理修正。工业品出厂价格指数的计算方法采用固定加权算术平均法。

根据价格传导规律,PPI 对 CPI 有一定的影响。PPI 反映生产环节价格水平,CPI 反映消费环节的价格水平。整体价格水平的波动一般首先出现在生产领域,然后通过产业链向下游产业扩散,最后波及消费品。所以,PPI 是衡量通货膨胀的先导性、潜在性指标,或者说,它是 CPI 出现波动的先声。一般认为,这种延迟的特性使得根据现在的 PPI 通货膨胀来粗略估计将来的 CPI 或通货膨胀率成为可能。但是,由于 CPI 不仅包括消费品价格,还包括服务价格,CPI 与 PPI 在统计口径上并非严格的对应关系,因此 CPI 与 PPI 的变化出现不一致的情况也是可能的。

四、股票价格指数

股票价格指数综合反映股票市场价格的变动程度,是反映证券市场行情变化的重要指标,不仅是广大证券投资者进行投资决策分析的依据,而且也被视为一个地区或国家宏观经济态势的“晴雨表”。股价指数的单位习惯上用“点”表示,即以基期为 100(或 1 000),每上升或下降 1 个单位称为 1 点。股价指数计算的方法很多,但一般以发行量为权数进行加权综合。其公式为

$$I = \sum p_{1i}q_i \Big/ \sum p_{0i}q_i \tag{7-22}$$

式中,p_{1i} 和 p_{0i} 分别为报告期和基期样本股的平均价格;q_i 为第 i 种股票的报告期发行量(也有采用基期的)。

世界各地的股票市场都有自己的股票价格指数。在一个国家里,同一股市往往有不

同的股票价格。下面介绍几种常见的股票价格指数。

（一）道琼斯股价平均数

道琼斯股价平均数(Dow Jones's average index)由美国的道琼斯公司计算并发布。自1884年第一次开始发布,迄今已有一个多世纪了。它是久负盛名、影响最广泛的一种股票价格指数。

道琼斯股价平均数以在纽约交易所挂牌上市交易的一些著名大公司的股票为编制对象。最初采用简单算术平均方法计算,将采样股票价格总额除以公司数,反映的是每一公司的平均股票价格总额。考虑到增资和折股等各种非市场因素对股票总股数的影响,后来采用除数修正法,即将各种采样股票价格总和除以一个修正后的除数来计算道琼斯股价平均数。除数修正公式为

$$修正后的新除数=\frac{非市场因素影响后的各种采样股票理论价格之和}{非市场因素影响前各种采样股票收盘价之和}\times 原先除数$$

$$道琼斯股价平均数=\frac{采样股票价格总和}{修正后的新除数}$$

人们通常引用的道琼斯股价指数实际是一组平均数,包括以下几类。

(1) 道琼斯工业股价平均数。由美国30家著名工商业公司股票组成采样股,主要用以反映整个工商业股票的价格水平。在许多场合,也被用作道琼斯股价平均数的代表。

(2) 交通运输业股价平均数。以美国20家著名的交通运输公司的股票为采样股,其中有8家铁路公司、8家航空公司和4家公路货运公司。

(3) 公用事业股价平均数。以美国15家最大公用事业公司的股票为采样股,反映公用事业类股票的价格水平。

(4) 股价综合平均数。以上述三种股价平均数所涉及的共65家公司的股票为采样股综合得到的股价平均数,反映整个股票市场价格的变化趋势。

（二）标准普尔指数

标准普尔指数是由美国最大的证券研究机构标准普尔(Standard&Poor)公司于1923年开始编制的,1957年起至今,标准普尔指数的采样股票一直保持在500种之多,包括400种工业股、20种运输股、40种金融股和40种公用事业股。该指数对比的基期为1941—1943年,是以基期各种股票发行量为权数,对所有采样股票价格采用拉氏公式计算而成的。由于它包括的股票市价总值约占纽约证券交易所上市股票的75%。因此,该指数具有较强的代表性和广泛的影响力,能较全面地反映股票市场价格的变动,在国际金融市场上影响也较大。

（三）香港恒生指数

1969年11月24日,香港恒生银行编制并首次公开发表香港恒生指数(Heng Seng index,HSI),它是香港证券市场上最有代表性的股票价格指数。

香港恒生指数共选择了33种具有代表性的股票(成分股)为指数计算对象。其中,金

融业 4 种,公用事业 6 种,地产业 9 种,其他行业 14 种。

香港恒生指数是以 1964 年 7 月 31 日为基期,基日指数定为 100。计算公式为

$$\text{即时指数} = \frac{\text{现时成分股的总市值}}{\text{上日收市时成分股的总市值}} \times \text{上日收市指数}$$

成分股的市值是按股价乘以发行股数计算的。因此,香港恒生指数也是以股票发行量为权数的加权综合指数。

(四) 上海证券交易所股价指数

上证综合指数是以 1990 年 12 月 19 日为基日(该日为上证所正式营业之日),基日定为 100,以所有在上海证券交易所上市的股票为编制范围,采用以股票发行量为权数的综合股价指数。计算公式为

$$\text{上证综合指数} = \frac{\text{报告期市价总值}}{\text{基日市价总值}} \times 100\%$$

式中,市价总值是股票市价乘发行股数;基日市价总值也称为除数。

当市价总值出现非交易因素(增股、配股、汇率等)变动时,原除数需修正,以维持指数的连续可比。修正公式为

$$\text{修正后的除数} = \frac{\text{修正后的市价总值}}{\text{修正前的市价总值}} \times \text{原除数}$$

(五) 深圳证券交易所股价指数

深证综合指数是以在深圳证券交易所上市的所有股票为对象编制的指数,1991 年 4 月 3 日为指数的基日,1991 年 4 月 4 日公布。深证综合指数是以发行量为权数,纳入指数计算范围的股票称为指数股。指数计算基本公式为

$$\text{指数} = \frac{\text{现时指数股总市值}}{\text{基日指数股总市值}} \times 100\%$$

若遇股市结构有所变动,其修正是用“连锁”方法计算得到的指数溯源于原有基期,以维持指数的连续性。每日连锁方法的计算公式为

$$\text{今日即时指数} = \frac{\text{今日即时指数股总市值}}{\text{经调整的上日指数股收市总市值}}$$

思　考　题

1. 简述指数的含义及分类。
2. 指出综合指数与平均指数的区别与联系。
3. 如何编制综合指数?
4. 指数体系的作用是什么?
5. 因素分析的内容和步骤有哪些?

练　习　题

1. 已给出某公司三种商品的销售量和单价数据，如表 7-9 所示。

表 7-9　某公司三种商品的销售量和单价资料

商品名称	计量单位	销售量		价格/元	
		2015 年	2016 年	2015 年	2016 年
甲	千克	8 000	8 800	10	10.5
乙	件	2 000	2 500	8	9
丙	盒	10 000	10 500	6	6.5

要求：(1) 计算三种商品的销售额总量指数。

(2) 以 2016 年销售量为权数计算三种商品的价格总指数。

(3) 以 2015 年单价为权数计算三种商品的销售量指数。

(4) 分析销售量和价格变动对销售额影响的绝对额和相对数。

2. 某家具公司生产三种产品的有关数据如表 7-10 所示。

表 7-10　某公司生产有关资料

产品名称	总生产费用/万元			报告期产量比基期增长/%
	基期	报告期	指数	
写字台	45.4	53.6	118.06	14
椅子	30	33.8	112.67	13.5
书柜	55.2	58.5	105.98	8.6

要求：

(1) 三种产品的生产费用总量指数。

(2) 以基期生产费用为权数的加权产量指数。

(3) 以报告期生产费用为权数的单位成本总指数。

(4) 分析产量和单位成本变动对总生产费用的影响。

3. 某厂生产两种产品，产量和原材料消耗的有关资料如表 7-11 所示。

表 7-11　某厂产量和原材料消耗的有关资料

产品名称	产量		每件原材料消耗量/千克		每千克原材料价格/元	
	基期	报告期	基期	报告期	基期	报告期
甲	20	22	5	4	50	58
乙	9	10	8	8	95	100

要求：分析原材料消耗总额的变动及各因素的影响程度和影响绝对额。

4. 某公司所属两个分厂的产品单位成本及生产量资料如表 7-12 所示，试分析该公司的产品总平均单位成本变动受分厂成本水平及产量结构变动的影响。

5. 给出某城市三个市场上有关同一种商品的销售资料，如表 7-13 所示。

表 7-12 某公司各分厂生产量与单位成本资料

分厂名称	单位成本/元		生产量/件	
	基期	报告期	基期	报告期
甲	10	9	300	1 300
乙	12	12.2	700	700
总厂	—	—	1 000	2 000

表 7-13 三市场商品的销售资料

市场名称	销售价格/(元/千克)		销售量/千克	
	基期	报告期	基期	报告期
A	2.5	3	740	560
B	2.4	2.8	670	710
C	2.2	2.4	550	820
合计	—	—	1 960	2 090

要求：

(1) 编制该商品总平均价格的可变构成指数、固定构成指数和结构变动影响指数。

(2) 建立指数体系，从相对数的角度进行总平均价格变动的因素分析。

(3) 综合分析销售总量变动和平均价格变动对该种商品销售总额的影响。

6. 已知某企业工资统计资料如表 7-14 所示。

表 7-14 某企业工资统计资料

指标分组	职工人数/人		工资总额/万元		平均工资/(元/人)		工资指数/%
	基期	报告期	基期	报告期	基期	报告期	
老职工	1 000	1 200	66.48	84.24	664.80	702.00	105.6
新职工	200	800	9.72	42.08	486.00	526.00	108.23
合计	1 200	2 000	76.2	126.32	635.00	631.60	99.46

从表 7-14 可以看出，新老职工报告期的平均工资分别比基期增长了 5.6%和8.23%，但全部职工的总平均工资反而下降了 0.54%。试计算分析说明原因。

7. 利用指数体系之间的关系回答下列问题。

(1) 某企业 2016 年同 2015 年相比，各种产品的产量增长了 8%，总生产费用增长了 12%，该企业 2016 年的单位成本有何变化？

(2) 某地区今年用同样多的人民币只能购买去年商品的 90%，求物价指数。若同样多的人民币比去年可多购买 10%的商品，物价指数是多少？

8. (1)某企业 2016 年比 2015 年职工人数增加了 2%，工业总产值增长了 17.3%。试计算该企业全员劳动生产率提高程度。

(2) 某市 2015 年社会商品零售额为 8 600 万元，2016 年增加为 12 890 万元，零售物价上涨了 11.5%。试计算该市零售额变动中零售量和价格两因素变动的影响程度和绝对值。

(3) 某产品生产总费用 2016 年为 12.9 万元，比 2015 年增加了 9 000 元，单位产品成本同比下降了 3%。试确定：①生产总费用指数；②产品物量指数；③由于成本降低而节约的费用额。

第八章

抽样推断

【统计知识 ABC】

[A]第二次世界大战中德军有多少辆坦克

第二次世界大战期间，盟军非常想知道德军总共制造了多少辆坦克。德国人在制造坦克时是墨守成规的，他们把坦克从 1 开始进行了连续编号。在战争过程中，盟军缴获了一些敌军坦克，并记录了它们的生产编号。那么怎样利用这些号码来估计坦克总数呢？在这个问题中，总体参数是未知的坦克总数 N，而缴获坦克的编号则是样本。

假设我们是盟军手下负责解决这个问题的统计人员，制造出来的坦克总数肯定大于等于记录的最大编号。为了找到它比最大编号大多少，我们先找到被缴获坦克编号的平均值，并认为这个值是全部编号的中点。因此样本均值乘以 2 就是总数的一个估计，当然要特别假设缴获的坦克代表了所有坦克的一个随机样本。这种估计 N 的公式的缺点是：不能保证均值的 2 倍一定大于记录中的最大编号。

N 的另一个点估计公式是：用观测到的最大编号乘以因子 $1+1/n$，其中 n 是被俘虏坦克个数。假如俘虏了 10 辆坦克，其中最大编号是 50，那么坦克总数的一个估计是 $(1+1/10)\times 50=55$。此处我们认为坦克的实际数略大于最大编号。

从战后发现的德军记录来看，盟军的估计值非常接近所生产的坦克的真实值。记录仍然表明统计估计比通常通过其他情报方式做出估计要大大接近于真实数目。统计学家们做得比间谍们更漂亮！

（资料来源：Gudmund R. Iversen，Mary Gergrn. 统计学——基本概念和方法[M]. 吴喜之，等，译. 北京：高等教育出版社；柏林：施普林格出版社，2000。）

[B]小概率事件

如果你是中国消费者协会的官员，负责治理缺斤少两的不法行为。假如你知道可口可乐公司，它们生产的一种瓶装雪碧，包装上标明其净含量是 500 毫升，在市场上随机抽取了 25 瓶，测得到其平均含量为 499.5 毫升，标准差为 2.63 毫升。你拿着这些数据可能做两件事，一是你做一个估计：该种包装的雪碧平均含量在 498.03～500.97 毫升，然后向消协写份报告；二是你做一个裁决：说“可口可乐公司有欺骗消费者的行为”的证据不足。前者是参数估计，后者是假设检验。

无论是假设检验还是参数估计，都要用到小概率。概率是从 0 到 1 的一个数，因此小概率就应该是接近 0 的一个数。著名的英国统计家 Ronald Fisher 把 1/20 作为标准，这

也就是 0.05,从此 0.05 或比 0.05 小的概率都被认为是小概率。Fisher 没有任何深奥的理由解释他为什么选择 0.05,只是说他忽然想起来的。小概率事件原理是,假定在一次实验中,小概率事件不发生,如果发生了,我们就有充足的理由拒绝原假设。

[C] 基尼系数

基尼系数(Gini coefficient),或译坚尼系数,是 20 世纪初意大利经济学家基尼根据洛伦茨曲线所定义的判断收入分配公平程度的指标。基尼系数是比例数值,介于 0 和 1 之间。越接近 0 就表明收入分配越是趋向平等;反之,收入分配越是趋向不平等。按照国际一般标准,0.4 以上的基尼系数表示收入差距较大,当基尼系数达到 0.6 以上时,则表示收入差距很大。国家统计局公布我国 2012 年基尼系数为 0.474,2010 年为 0.481。

设实际收入分配曲线和收入分配绝对平等曲线之间的面积为 A,实际收入分配曲线右下方的面积为 B,并以 A 除以 $(A+B)$ 的商表示不平等程度。这个数值被称为基尼系数或称洛伦茨系数。如果 A 为零,基尼系数为零,表示收入分配完全平等;如果 B 为零则系数为 1,收入分配绝对不平等。收入分配越是趋向平等,洛伦茨曲线的弧度越小,基尼系数也越小;反之,收入分配越是趋向不平等,洛伦茨曲线的弧度越大,那么基尼系数也越大。

(资料来源:http://baike.baidu.com/view/186.html.)

按照随机原则,以一定概率从总体中抽取一定容量的单位作为样本进行调查,根据样本统计量(描述样本数量特征指标)对总体参数(描述总体数量特征指标)做出具有一定可靠程度的估计与推断。这一完整过程,就是我们通常所说的抽样、抽样调查或抽样推断。抽样推断有两种类型:一类是参数估计,由对部分进行观测取得的数据对研究对象整体的数量特征取值给出估计方法;另一类是假设检验,由对部分进行观测取得的数据对研究对象的数量规律性是否具有某种指定特征进行检验。

第一节　参 数 估 计

一、参数估计概述

在许多实际问题中,总体被理解为我们所研究的那个统计指标,它在一定范围内取数值,而且是以一定的概率取各种数值的,从而形成一个概率分布,但是这个概率分布往往是未知的。例如,为了制定绿色食品的有关规定,我们需要研究蔬菜中残留农药的分布状况,对这个分布我们知之甚少,以致它属于何种类型我们都不清楚。有时我们可以断定分布的类型,如在农民收入调查中,根据实际经验和理论分析,如概率论中的中心极限定理,我们断定收入服从正态分布,但分布中的参数取何值却是未知的。这就导致了统计估计问题。统计估计问题专门研究由样本估计总体的未知分布或分布中的未知参数。直接对总体的未知分布进行估计的问题称为非参数估计;当总体分布类型已知,仅需对分布的未知参数进行估计的问题称为参数估计。本节我们研究参数估计问题。本节及以后假定抽样方法为放回简单随机抽样,样本的每个分量都与总体同分布,它们之间相互独立。

二、参数估计的基本方法

（一）估计量与估计值

（1）参数估计就是用样本统计量去估计总体参数。

（2）用来估计总体参数的统计量的名称称为估计量，如样本均值、样本比例、样本方差等都可以是一个估计量。

（3）估计量的具体数值称为估计值。

（二）点估计与区间估计

参数估计方法有点估计与区间估计两种方法。

1．参数估计的点估计法

设总体 X 的分布类型已知，但包含有未知参数 θ，从总体中抽取一个简单随机样本 $(X_1,X_2,\cdots,X_n)$，欲利用样本提供的信息对总体未知参数 θ 进行估计。构造一个适当的统计量

$$\hat{\theta}=T(X_1,X_2,\cdots,X_n)$$

作为 θ 的估计，称 $\hat{\theta}$ 为未知参数 θ 的点估计量（point estimate）。当有了一个具体的样本观察值 $(x_1,x_2,\cdots,x_n)$ 后，将其代入估计量中就得到估计量的一个具体观察值 $T(x_1,x_2,\cdots,x_n)$，称为参数θ 的一个点估计值。今后点估计量和点估计值这两个名词将不强调它们的区别，统称为点估计，根据上下文不难知道此处的点估计究竟是点估计量还是点估计值。

通俗地说，用样本估计量的值直接作为总体参数的估计值称为点估计。

2．参数估计的区间估计法

在参数估计中，虽然点估计可以给出未知参数的一个估计，但不能给出估计的精度。为此人们希望利用样本给出一个范围，要求它以足够大的概率包含待估参数真值。这就导致了区间估计（interval estimation）问题。

设 θ 是未知参数，$(X_1,X_2,\cdots,X_n)$ 是来自总体的样本，构造两个统计量 $\hat{\theta}_1=T_1(X_1,X_2,\cdots,X_n)$，$\hat{\theta}_2=T_2(X_1,X_2,\cdots,X_n)$，对于给定的 $\alpha(0<\alpha<1)$，若 $\hat{\theta}_1$、$\hat{\theta}_1$ 满足

$$P\{\hat{\theta}_1\leqslant\theta\leqslant\hat{\theta}_2\}=1-\alpha$$

则称随机区间 $[\hat{\theta}_1,\hat{\theta}_2]$ 是参数 θ 的置信水平（confidence level）为 $(1-\alpha)$ 的置信区间（confidence interval），$(1-\alpha)$ 称为 $[\hat{\theta}_1$、$\hat{\theta}_2]$ 的置信系数，$\hat{\theta}_1$、$\hat{\theta}_2$ 称为置信限（confidence limit）。

这里有几点需要说明。

（1）区间 $[\hat{\theta}_1$、$\hat{\theta}_2]$ 的端点 $\hat{\theta}_1$、$\hat{\theta}_2$ 及长度 $(\hat{\theta}_2-\hat{\theta}_1)$ 都是样本的函数，从而都是随机变量，因此 $[\hat{\theta}_1,\hat{\theta}_2]$ 是一个随机区间。

（2）$P\{\hat{\theta}_1\leqslant\theta\leqslant\hat{\theta}_2\}=1-\alpha$ 是说随机区间 $[\hat{\theta}_1,\hat{\theta}_2]$ 以 $(1-\alpha)$ 的概率包含未知参数真值，

区间长度$(\hat{\theta}_2-\hat{\theta}_1)$描述估计的精度,置信水平$(1-\alpha)$描述了估计的可靠度。

(3) 因为未知参数θ是非随机变量,所以不能说θ落入区间$[\hat{\theta}_1,\hat{\theta}_2]$的概率是$(1-\alpha)$,而应是随机区间$[\hat{\theta}_1,\hat{\theta}_2]$包含$\theta$的概率是$(1-\alpha)$。

通俗地说,在点估计的基础上,给出总体参数的一个范围称为区间估计。

三、总体均值的区间估计

(一) 正态总体且方差已知;或非正态总体、方差未知、大样本情况下

在这种情况下,样本均值的抽样分布呈正态分布,其数学期望为总体均值μ,方差为$\frac{\sigma^2}{n}$。则$\left(\bar{X}\pm Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)$称为总体均值在$(1-\alpha)$置信水平下的置信区间。

设样本$(X_1,X_2,\cdots,X_n)$来自正态总体$N(\mu,\sigma^2)$,μ是总体均值,σ^2是总体方差,当σ^2已知时,数理统计证明$\bar{X}$服从正态分布$N\left(\mu,\frac{\sigma^2}{n}\right)$,从而$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$服从标准正态分布$N(0,1)$,对给定的置信系数$1-\alpha$查$N(0,1)$表可得上$\frac{\alpha}{2}$分位点$Z_{\frac{\alpha}{2}}$,使得

$$P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\leqslant Z_{\frac{\alpha}{2}}\right\}=1-\alpha$$

从而有

$$P\left\{\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leqslant\mu\leqslant\bar{X}+Z_{\frac{\alpha}{2}}\frac{\alpha}{\sqrt{n}}\right\}=1-\alpha$$

取

$$\hat{\mu}_1=\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\quad\hat{\mu}_2=\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\tag{8-1}$$

则$[\hat{\mu}_1,\hat{\mu}_2]$即是$\mu$的置信水平为$(1-\alpha)$的置信区间。

【例 8-1】 保险公司从投保人中随机抽取 36 人,计算得 36 人的平均年龄$\bar{X}=39.5$岁,已知投保人平均年龄近似服从正态分布,标准差为 7.2 岁,试求全体投保人平均年龄的置信水平为 99%的置信区间。

解 $1-\alpha=0.99$, $\alpha=0.01$, $\frac{\alpha}{2}=0.005$, $1-\frac{\alpha}{2}=0.995$

使用 NORMSINV 函数得:$\left[\text{注:在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.005}=2.575\,829\,304$$

$$\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}=39.5-2.575\,829\,304\times\frac{7.2}{\sqrt{36}}=36.41$$

$$\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}=39.5+2.575\,829\,304\times\frac{7.2}{\sqrt{36}}=42.59$$

故全体投保人平均年龄的置信水平为 99%的置信区间为[36.41,42.59]

在不重复抽样条件下,置信区间为

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{8-2}$$

【例 8-2】 一家食品公司，每天大约生产袋装食品若干，按规定每袋的重量应为 100 克。为对产品质量进行检测，该企业质检部门采用抽样技术，每天抽取一定数量的食品，以分析每袋重量是否符合质量要求。现从某一天生产的一批食品 8 000 袋中随机抽取了 25 袋(不重复抽样)，测得它们的重量如图 8-1 所示。

	A	B	C	D	B
1	112.5	101	103	102	100.5
2	102.6	107.5	95	108.8	115.6
3	100	123.5	102	101.6	102.2
4	116.6	95.4	97.8	108.6	105
5	136.8	102.8	101.5	98.4	93.3

图 8-1　25 袋食品重量资料

已知产品重量服从正态分布，且总体方差为 100 克。试估计该批产品平均重量的置信区间，置信水平为 95%。

解　已知 $\sigma^2=100$ 克，$n=25$，$1-\alpha=0.95$，$\frac{\alpha}{2}=0.025$，$1-\frac{\alpha}{2}=0.975$

使用 NORMSINV 函数得：$\left[\text{注：在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.025}=1.959\ 963\ 985$$

根据样本资料，计算的样本均值为(也可直接通过 Excel 软件中的描述统计功能计算)：

$$\overline{X}=\frac{\sum x}{n}=\frac{2\ 634}{25}=105.36$$

根据式(8-2)得

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 105.36 \pm 1.96 \times \sqrt{\frac{100}{25}} \times \sqrt{\frac{8\ 000-25}{8\ 000-1}}$$

即 105.36±3.914 115=(101.445 9，109.274 1)，该批产品平均重量在 95%置信水平下的置信区间为 101.445 9～109.274 1。

若总体方差 σ^2 未知，可用样本方差 S^2 代替。

【例 8-3】 承例 8-1，假定保险公司从投保人中随机抽取 36 人，得到他们的年龄数据如图 8-2 所示。

	A	B	C	D	E	F
1	23	35	39	27	36	44
2	36	42	46	43	31	33
3	42	53	45	54	47	24
4	34	28	39	36	44	40
5	39	49	38	34	48	50
6	34	39	45	48	45	32

图 8-2　36 名投保人年龄资料

若总体方差未知，试建立投保人年龄 90% 的置信区间。

解　已知 $n=36$，$1-\alpha=90\%$，$1-\frac{\alpha}{2}=0.95$

使用 NORMSINV 函数得：$\left[\text{注：在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.025}=1.644\ 853\ 627$$

由于总体方差 σ^2 未知，但为大样本，故可用样本方差代替。

根据样本资料计算的样本均值和样本标准差为

$$\overline{X}=\frac{\sum x}{n}=\frac{1\ 422}{36}=39.50$$

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=7.773\ 582$$

(样本均值和样本标准差的计算,也可直接通过 Excel 2013 中的描述统计功能计算,计算结果如图 8-3 所示)。

则置信区间为

$$\overline{X}\pm Z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}=39.5\pm1.644\ 853\ 627\times\frac{7.773\ 582}{\sqrt{36}}$$

即 39.5±2.13=(37.37,41.63),投保人平均年龄在 90%的置信水平下的置信区间为 37.37~41.63 岁。



	A	B	C	D	E	F
1	23		列1			
2	36					
3	42		平均	39.5		
4	34		标准误差	1.2956		
5	39		中位数	39		
6	34		众数	39		
7	35		标准差	7.77358		
8	42		方差	60.4286		
9	53		峰度	-0.4476		
10	28		偏度	-0.2266		
11	49		区域	31		
12	39		最小值	23		
13	39		最大值	54		
14	46		求和	1422		
15	45		观测数	36		
16	39					

图 8-3 描述统计运行结果

(二)正态总体、方差未知、小样本情况下

如果总体服从正态分布,无论样本容量大小,样本均值的抽样分布都服从正态分布。只要总体方差已知,即使在小样本情况下,也可以按式(8-1)或式(8-2)计算总体均值的置信区间。如果总体方差 σ^2 未知,需用样本方差 s^2 代替,在小样本情况下,应用 t 分布来建立总体均值的置信区间。

t 分布是类似正态分布的一种对称分布,它通常要比正态分布平坦和分散。随着自由度的增大,t 分布逐渐趋于正态分布。

正态总体、方差未知、小样本情况下,总体均值在(1−α)置信水平下的置信区间为

$$\overline{X}\pm t_{\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}\quad(\text{重复抽样条件下})\tag{8-3}$$

$$\overline{X}\pm t_{\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\quad(\text{不重复抽样条件下})\tag{8-4}$$

式中,$t_{\frac{\alpha}{2}}(n-1)$为自由度为$(n-1)$的 t 分布临界值,可以查 t 分布临界值表得到,也可由 Excel 计算得到。

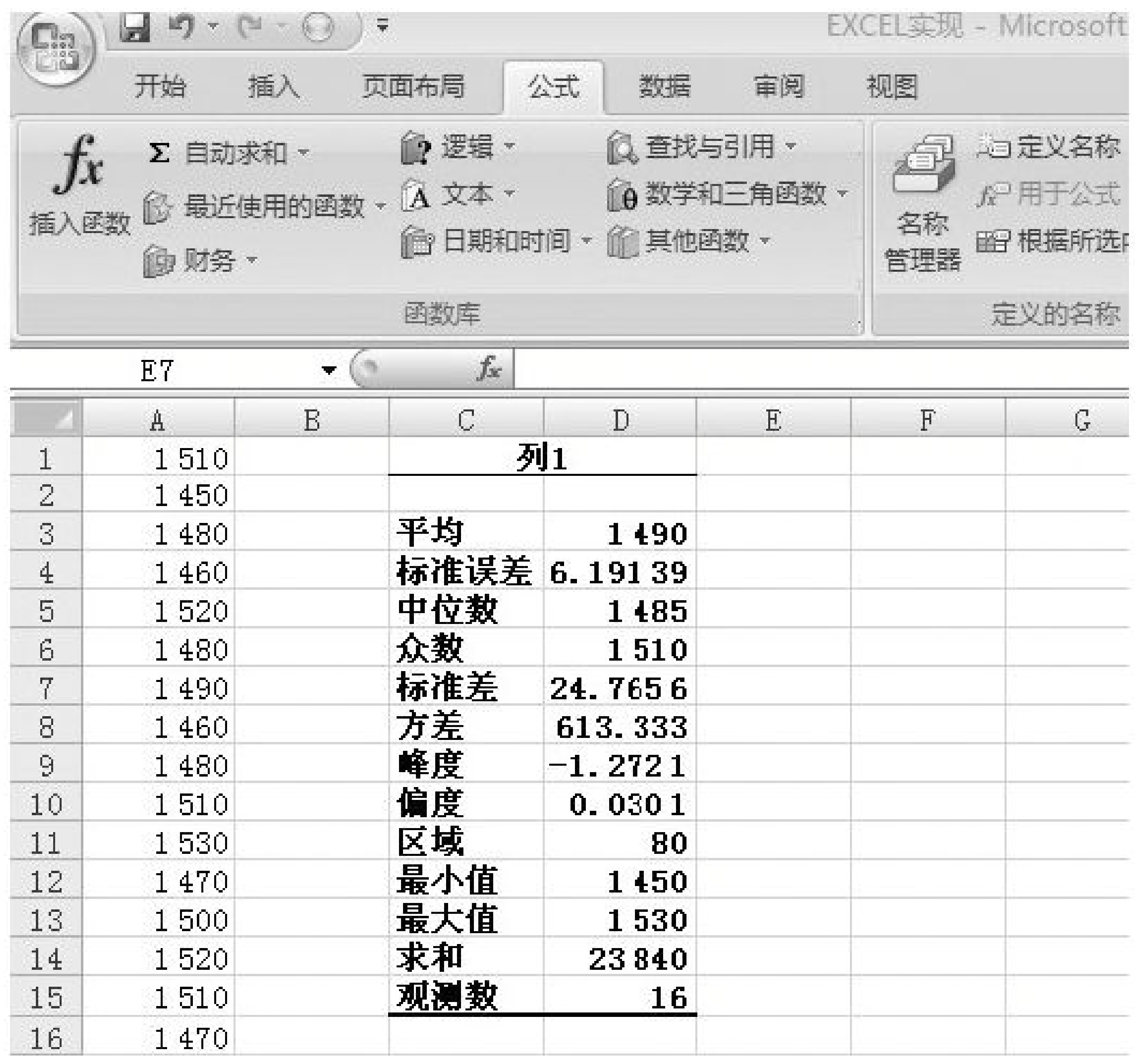
Excel 计算,可使用粘贴函数“TINV”完成。操作步骤依次为:TINV→$\frac{\alpha}{2}$→$(n-1)$→确定。

【例 8-4】 已知某种电子元件的寿命服从正态分布,现从一批电子元件中随机抽取 16 只,测得其寿命如图 8-4 中的 A 列所示。试建立该批电子元件使用寿命 95%的置信区间。

根据样本资料计算的样本均值和样本标准差为

$$\overline{X}=\frac{\sum x}{n}=\frac{23\ 840}{16}=1\ 490,\quad s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=24.765\ 57$$

(样本均值和样本标准差的计算,也可直接通过 Excel 软件中的描述统计功能计算,



	A	B	C	D	E	F	G
1	1 510		列1				
2	1 450						
3	1 480		平均	1 490			
4	1 460		标准误差	6.191 39			
5	1 520		中位数	1 485			
6	1 480		众数	1 510			
7	1 490		标准差	24.765 6			
8	1 460		方差	613.333			
9	1 480		峰度	-1.272 1			
10	1 510		偏度	0.030 1			
11	1 530		区域	80			
12	1 470		最小值	1 450			
13	1 500		最大值	1 530			
14	1 520		求和	23 840			
15	1 510		观测数	16			
16	1 470						

图 8-4　16 只电子元件寿命原始数据及描述统计部分结果

计算结果如图 8-4 所示)。

已知 $1-\alpha=95\%$，$\frac{\alpha}{2}=0.025$，$n=16$

使用 TINV 函数得：$\left(\text{注：在 probability 栏内输入}\frac{\alpha}{2}\text{值}\right)$

$$t_{0.025}(15)=2.489\ 879\ 694$$

则该批电子元件平均使用寿命 95%的置信区间为

$$\overline{X}\pm t_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}=1\ 490\pm 2.489\ 879\ 694\times\frac{24.77}{\sqrt{16}}$$

即 $1\ 490\pm15.42=(1\ 474.58,1\ 505.42)$，该批电子元件平均使用寿命在 95%的置信水平下的置信区间为 1 474.58～1505.42 小时。

现将总体均值的区间估计总结如图 8-5 所示。



总体分布	样本容量	σ 已知	σ 未知
正态分布	大样本	$\bar{x}\pm Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$	$\bar{x}\pm Z_{\frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}$
	小样本		$\bar{x}\pm t_{\frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}$
非正态分布	大样本		$\bar{x}\pm Z_{\frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}$

图 8-5　不同情况下总体均值的区间估计

四、总体比例的区间估计

在大样本[一般经验规则：$np\geqslant5$ 和 $n(1-p)\geqslant5$]条件下，样本比例的抽样分布可用正态分布近似。在这种情况下，数理统计已经证明如下结论。

置信水平为$(1-\alpha)$的置信区间为

$$p\pm Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{p(1-p)}{n}}\quad(\text{重复抽样})\qquad(8\text{-}5)$$

$$p\pm Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}\quad(\text{不重复抽样})\qquad(8\text{-}6)$$

【例 8-5】 某城市想要估计下岗职工中女性所占的比例，采取重复抽样方法随机抽取了 100 名下岗职工，其中 65 人为女性。试以 95%的置信水平估计该城市下岗职工中

女性所占比例的置信区间。

解 已知 $n=100$，$p=\frac{65}{100}=65\%$，$1-\alpha=0.95$，$1-\frac{\alpha}{2}=0.975$

使用 NORMSINV 函数得：$\left[\text{注：在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.025}=1.959\,963\,985$$

$$p\pm Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{p(1-p)}{n}}=65\%\pm1.959\,963\,985\times\sqrt{\frac{65\%\times(1-65\%)}{100}}$$

即 65%±9.35%=(55.65%,74.35%)，95%的置信水平下估计该城市下岗职工中女性所占比例的置信区间为 55.65%～74.35%。

【例 8-6】 某企业共有职工 1 000 人，企业准备实行一项改革，在职工中征求意见，采用不重复抽样方法，随机抽取 200 人作为样本，调查结果显示，有 150 人表示赞成这项改革，有 50 人表示反对。试以 95%的置信水平确定赞成改革的人数比例的置信区间。

解 已知 $n=200$，$p=\frac{150}{200}=75\%$，$1-\alpha=0.95$，$1-\frac{\alpha}{2}=0.975$

使用 NORMSINV 函数得：$\left[\text{注：在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.025}=1.959\,963\,985$$

$$p\pm Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$

$$=75\%\pm1.959\,963\,985\times\sqrt{\frac{75\%(1-75\%)}{200}\left(\frac{1\,000-200}{1\,000-1}\right)}$$

即 75%±5.37%=(69.63%,80.37%)，95%的置信水平下估计赞成改革的人数比例的置信区间为 69.63%～80.37%。

五、总体方差的区间估计

数理统计证明，对于容量为 n 的正态总体样本方差 s^2，若总体方差为 σ^2，则 $\frac{(n-1)s^2}{\sigma^2}$ 服从自由度为 $(n-1)$ 的 χ^2 的分布。对给定的置信水平 $(1-\alpha)$，查 χ^2 分布表可得 $\frac{\alpha}{2}$ 分位点 $\chi^2_{\frac{\alpha}{2}}(n-1)$ 和 $\left(1-\frac{\alpha}{2}\right)$ 分位点 $\chi^2_{1-\frac{\alpha}{2}}(n-1)$，使得

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(n-1)\leqslant\frac{(n-1)s^2}{\sigma^2}\leqslant\chi^2_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

从而有

$$P\left\{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\leqslant\sigma^2\leqslant\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right\}=1-\alpha$$

取

$$\hat{\sigma}_1^2=\frac{(n-1)s^2}{X^2_{\frac{\alpha}{2}}(n-1)}，\quad\hat{\sigma}_2^2=\frac{(n-1)s^2}{X^2_{1-\frac{\alpha}{2}}(n-1)}$$

则$[\hat{\sigma}_1^2, \hat{\sigma}_2^2]$即是$\sigma^2$的置信水平为$(1-\alpha)$的置信区间。即

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \tag{8-7}$$

【例 8-7】 某食品厂从生产的罐头中随机抽取 15 个称其重量，得样本方差 $s^2 = 1.65^2$，设罐头重量服从正态分布，试求其方差的置信水平为 90%的置信区间。

解　$n=15$，$1-\alpha=90\%$，$\frac{\alpha}{2}=0.05$，$1-\frac{\alpha}{2}=0.95$

使用 CHIINV 函数得[函数格式：CHIINV(0.05,14)；CHIINV(0.95,14)]

$$\chi^2_{0.05}(14) = 23.685,\quad \chi^2_{0.95}(14) = 6.571$$

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} = \frac{14\times 1.65^2}{23.685} = 1.61,\quad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} = \frac{14\times 1.65^2}{6.571} = 5.8$$

故总体方差的置信水平为 90%的置信区间为[1.61,5.8]。

第二节　假设检验

一、假设检验的有关概念

（一）统计假设

统计假设是对总体参数的具体数值所做的陈述。总体参数包括总体均值μ、比例π、方差σ^2等。

（二）假设检验

假设检验是先对总体的参数(或分布形式)提出某种假设，然后利用样本信息判断假设是否成立的过程。有参数检验和非参数检验两种。逻辑上运用反证法，统计上依据小概率原理。

（三）原假设

原假设是研究者想收集证据予以反对的假设，又称“零假设”，用符号表示为H_0。

之所以用零来修饰原假设，其原因是原假设的内容总是没有差异或没有改变，或变量间没有关系等。关于样本统计量如样本均值或样本均值之差的零假设是没有意义的，因为样本统计量是已知的，当然能说出它们等于几或是否相等。

（四）备择假设

备择假设也称“研究假设”，是研究者想收集证据予以支持的假设，用符号表示为H_1。

（五）双侧检验与单侧检验

如果备择假设没有特定的方向性，并含有符号“$\neq$”，这样的称为双侧检验或双尾检验。

如果备择假设具有特定的方向性,并含有符号“>”或“<”的假设检验,称为单侧检验或单尾检验。备择假设的方向为“<”,称为左侧检验;备择假设的方向为“>”,称为右侧检验。

(六) 提出假设

原假设和备择假设是一个完备事件组,而且相互对立。在一项假设检验中,原假设和备择假设必有一个成立,而且只有一个成立。

先确定备择假设,再确定原假设。

等号“=”总是放在原假设上。

因研究目的不同,对同一问题可能提出不同的假设(也可能得出不同的结论)。

【例 8-8】 一种零件的生产标准是直径应为 10 厘米,为对生产过程进行控制,质量监测人员定期对一台加工机床检查,确定这台机床生产的零件是否符合标准要求。如果零件的平均直径大于或小于 10 厘米,则表明生产过程不正常,必须进行调整。如果检验生产过程是否正常,原假设和备择假设分别是什么?

解 研究者想收集证据予以证明的假设应该是“生产过程不正常”。

建立的原假设和备择假设为

$$H_0: \mu = 10 \text{ 厘米}, \quad H_1: \mu \neq 10 \text{ 厘米}$$

【例 8-9】 某品牌洗涤剂在它的产品说明书中声称:平均净含量不少于 500 克。从消费者的利益出发,有关研究人员要通过抽检其中的一批产品来验证该产品制造商的说明是否属实。试陈述用于检验的原假设与备择假设。

解 研究者抽检的意图是倾向于证实这种洗涤剂的平均净含量并不符合说明书中的陈述。

建立的原假设和备择假设为

$$H_0: \mu \geqslant 500, \quad H_1: \mu < 500$$

【例 8-10】 一家研究机构估计,某城市中家庭拥有汽车的比例超过 30%。为验证这一估计是否正确,该研究机构随机抽取了一个样本进行检验。试陈述用于检验的原假设与备择假设。

解 研究者想收集证据予以支持的假设是“该城市中家庭拥有汽车的比例超过 30%”。

建立的原假设和备择假设为

$$H_0: \mu \leqslant 30\%, \quad H_1: \mu > 30\%$$

二、假设检验中的两类错误

(一) 第Ⅰ类错误(弃真错误)

原假设为真时拒绝原假设。第Ⅰ类错误的概率记为 α(alpha),被称为显著性水平。

(二) 第Ⅱ类错误(取伪错误)

原假设为假时未拒绝原假设。第Ⅱ类错误的概率记为 β(Beta)。

在这两类错误中，相对更加严重的是第Ⅰ类错误，所以 α 的取值应尽可能小。

（三）显著性水平α

显著性水平是一个概率值。原假设为真时，拒绝原假设的概率，被称为抽样分布的拒绝域，表示为 α，常用的 α 值有 0.01、0.05、0.10，由研究者事先确定。

我们可以在事先确定用于拒绝原假设 H_0 的证据必须强到何种程度。这等于说我们要求多小的 P 值，而这个 P 值就叫显著性水平。显著性水平表示总体中某一类数据出现的经常程度。

假如我们选择 $\alpha=0.05$，样本数据能拒绝原假设的证据要强到：当 H_0 正确时，这种样本结果发生的频率不超过 5%；如果我们选择 $\alpha=0.01$，就是要求拒绝 H_0 的证据要更强，这种样本结果发生的频率只有 1%。

如果 P 值小于或等于 α，我们称该组数据不利于原假设的证据有 α 的显著性水平。

（四）统计显著性（significant）

significant（显著的）一词的意义在这里并不是“重要的”，而是指“非偶然的”。

在假设检验中，如果样本提供的证据拒绝原假设，我们说检验的结果是显著的，如果不拒绝原假设，我们则说结果是不显著的。

一项检验在统计上是“显著的”，意思是指：这样的（样本）结果不是偶然得到的，或者说，不是靠机遇能够得到的。

拒绝原假设，表示这样的样本结果并不是偶然得到的；不拒绝原假设（拒绝原假设的证据不充分），则表示这样的样本结果只是偶然得到的。

（五）假设检验中的小概率原理

在一次试验中，一个几乎不可能发生的事件发生的概率称为小概率；在一次试验中小概率事件一旦发生，我们就有理由拒绝原假设；小概率由研究者事先确定。

概率是 0～1 的一个数，因此小概率就应该是接近 0 的一个数。著名的英国统计家 Ronald Fisher 把 1/20 作为标准，这也就是 0.05，从此 0.05 或比 0.05 小的概率都被认为是小概率。Fisher 没有任何深奥的理由来解释他为什么选择 0.05，只是说他忽然想起来的。

三、检验统计量与拒绝域

（一）检验统计量

检验统计量是根据样本观测结果计算得到的，并据以对原假设和备择假设做出决策的某个样本统计量。是对样本估计量的标准化结果，即原假设 H_0 为真时点估计量的抽样分布。

$$\text{标准化检验统计量}=\frac{\text{点估计量}-\text{假设值}}{\text{点估计量的抽样标准差}} \tag{8-8}$$

(二) 双侧检验的拒绝域

双侧检验的拒绝域如图 8-6 所示。

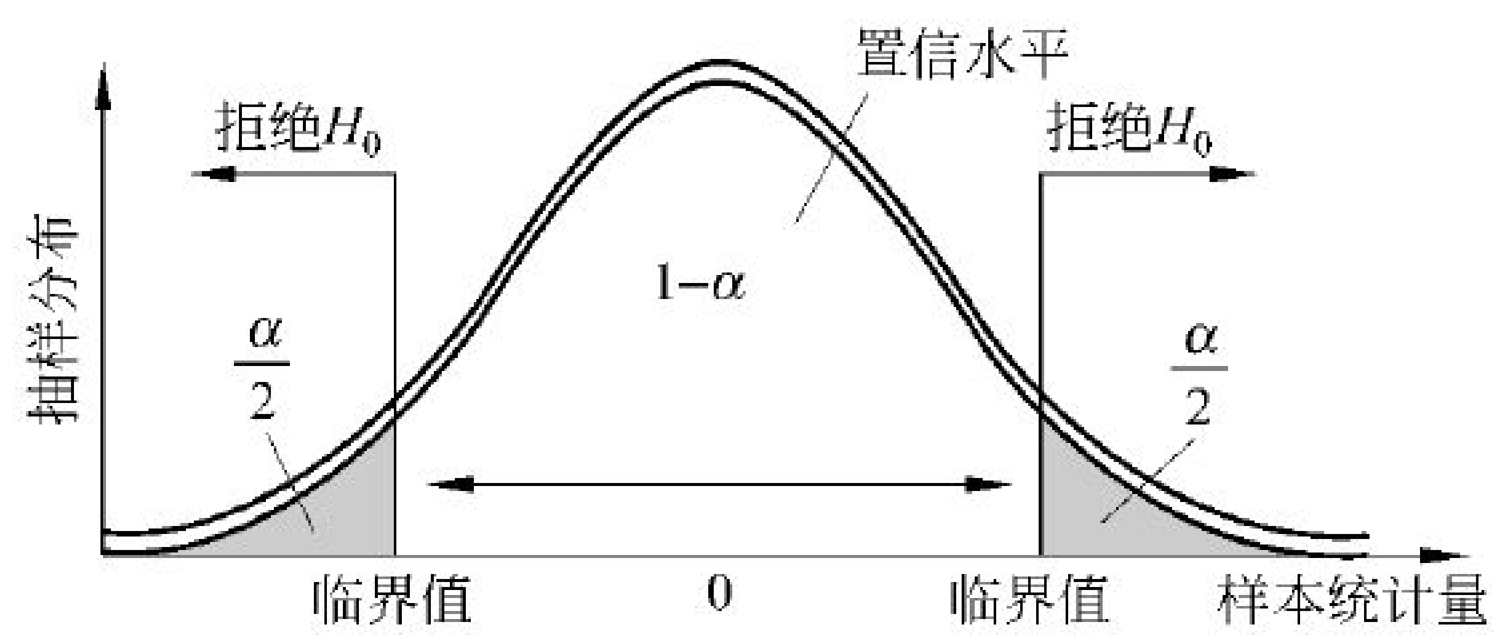


图 8-6 双侧检验的拒绝域

(三) 左侧检验的拒绝域

左侧检验的拒绝域如图 8-7 所示。

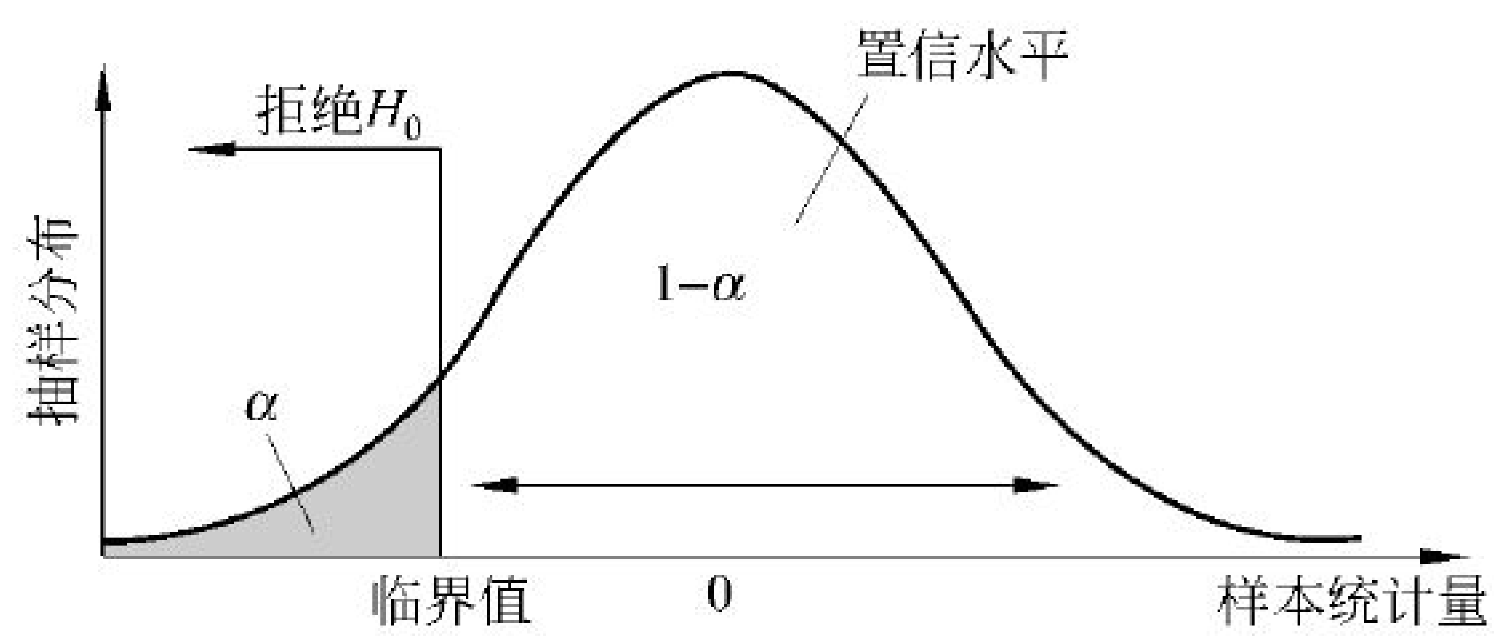


图 8-7 左侧检验的拒绝域

(四) 右侧检验的拒绝域

右侧检验的拒绝域如图 8-8 所示。

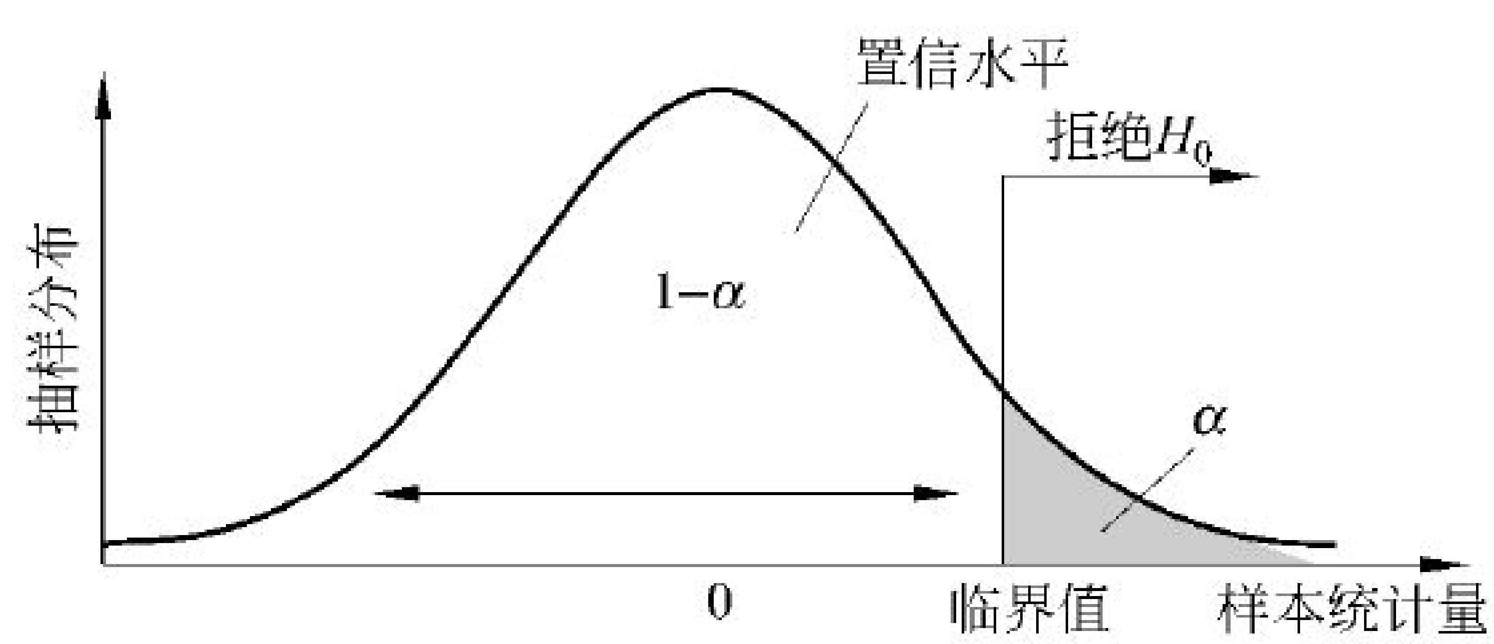


图 8-8 右侧检验的拒绝域

(五) 决策规则

(1) 给定显著性水平 α,查表得出相应的临界值 Z_α 或 $Z_{\frac{\alpha}{2}}$, t_α 或 $t_{\frac{\alpha}{2}}$。

(2) 将检验统计量的值与 α 水平的临界值进行比较做出决策。

① 双侧检验:统计量>临界值,拒绝 H_0。

② 左侧检验:统计量<临界值,拒绝 H_0。

③ 右侧检验:统计量>临界值,拒绝 H_0。

四、利用 *P* 值进行决策

P 值(*P*-value)是指如果原假设为真,所得到的样本结果会像实际观测结果那样极端或更极端的概率;*P* 值是反映实际观测到的数据与原假设 H_0 之间不一致的程度;被称为观察到的(或实测的)显著性水平;决策规则:若 p 值小于 α, 拒绝 H_0。

P 值告诉我们:如果原假设是正确的话,我们的观测数据就不可能得到。不可能得到的数据,就是原假设不对的合理证据。

五、假设检验的步骤

假设检验的步骤如下。

(1) 陈述原假设和备择假设。

(2) 从所研究的总体中抽出一个随机样本。

(3) 确定一个适当的检验统计量,并利用样本数据算出其具体数值。

(4) 确定一个适当的显著性水平,并计算出其临界值,指定拒绝域。

(5) 将统计量的值与临界值进行比较,做出决策。

统计量的值落在拒绝域,拒绝 H_0,否则不能拒绝 H_0。也可以直接利用 *P* 值做出决策,若 p 值小于 α, 拒绝 H_0。

六、总体均值的假设检验

(一) 正态总体或非正态总体大样本($n>30$)情况下,使用 z 检验统计量

(1) σ^2 已知,总体均值检验的统计量为

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \tag{8-9}$$

【例 8-11】 一种罐装饮料采用自动生产线生产,每罐的容量是 255 毫升,标准差为 5 毫升。为检验每罐容量是否符合要求,质检人员在某天生产的饮料中随机抽取了 40 罐进行检验,测得每罐平均容量为 255.8 毫升。取显著性水平 $\alpha=0.05$,检验该天生产的饮料容量是否符合标准要求?

解　由已知得:$\mu_0=255$, $\sigma=5$, $n=40$, $\bar{x}=255.8$, $\alpha=0.05$

$$H_0: \mu=255; \quad H_1: \mu\neq 255$$

计算统计量的具体数据,有

$$z=\frac{255.8-255}{5/\sqrt{40}}=1.01$$

根据给定的显著性水平 $\alpha=0.05$,查标准正态分布表,得 $z_{\frac{\alpha}{2}}=z_{0.025}=1.96$。也可以利用 Excel 中的统计函数 normsiv$\left(1-\frac{\alpha}{2}\right)$计算得到。

由于$|z|=1.01<z_{\frac{\alpha}{2}}=1.96$,所以,不拒绝原假设。检验结果表明:样本提供的证据不足以推翻原假设,说明该天生产的饮料符合标准要求。

(2) σ^2 未知,可以用样本方差 s^2 来近似代替总体方差,总体均值检验统计量为

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{8-10}$$

【例 8-12】 一种机床加工的零件尺寸绝对平均误差允许值为 1.35 毫米。生产厂家现采用一种新的机床进行加工以期进一步降低误差。为检验新机床加工的零件平均误差与旧机床相比是否有显著降低,从某天生产的零件中随机抽取 50 个进行检验。50 个零件的平均数为 1.215 2,样本标准差是 0.365 749。利用以上数据,检验新机床加工的零件尺寸的平均误差与旧机床相比是否有显著降低?($\alpha=0.01$)

解 $H_0:\mu\geqslant 1.35$, $H_1:\mu<1.35$

计算检验统计量的具体数值

$$z=\frac{1.215\,2-1.35}{0.365\,749/\sqrt{50}}=-2.606\,1$$

根据给定的显著性水平 $\alpha=0.01$,查标准正态分布得 $z_\alpha=z_{0.01}=-2.33$,也可以利用 Excel 中的统计函数 normsiv($1-\alpha$)计算得到。

由于 $z=-2.606\,1<z_{0.01}=-2.33$,所以拒绝原假设。检验结果表明:新机床加工的零件尺寸的平均误差与旧机床相比有显著降低。

(二)小样本情况下,依照总体方差是否已知来选择合适的统计量

(1) σ^2 已知:即使是在小样本情况下,检验统计量仍服从标准正态分布。检验统计量和检验程序与大样本时完全相同,此处不再赘述。

(2) σ^2 未知,小样本:需要用样本方差 s^2 代替总体方差 σ^2,此时检验统计量服从自由度为($n-1$)的 t 分布,检验统计量为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{8-11}$$

表 8-1 所示为小样本情况下一个总体均值的检验方法。

表 8-1 小样本情况下一个总体均值的检验方法

	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\mu=\mu_0;H_1:\mu\neq\mu_0$	$H_0:\mu\geqslant\mu_0;H_1:\mu<\mu_0$	$H_0:\mu\leqslant\mu_0;H_1:\mu>\mu_0$
检验统计量	σ 未知:$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ $\left(\sigma\text{ 已知}:z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$		
α 与拒绝域	$\|t\|>t_{\frac{\alpha}{2}}(n-1)$	$t<-t_\alpha(n-1)$	$t>t_\alpha(n-1)$
P 值决策准则	$P<\alpha$,拒绝 H_0		

【例 8-13】 一种汽车配件的平均长度要求为 12 厘米,高于或低于该标准均被认为是不合格的。汽车生产企业在购进配件时,通常是经过招标,然后对中标的配件提供商提供的样品进行检验,以决定是否购进。现对一个配件提供商的 10 个样本进行检验,结果如下(单位:cm):

12.2　10.8　12.0　11.8　11.9　12.4　11.3　12.2　12.0　12.3

假定该供货商生产的配件长度服从正态分布，在 0.05 的显著性水平下，检验该供货商提供的配件是否符合要求？

解　$H_0: \mu=12, H_1: \mu\neq 12$

依据样本数据计算得：$\bar{x}=11.89, s=0.4932$

由于是小样本，所以采用 t 统计量

$$t=\frac{11.89-12}{0.4932/\sqrt{10}}=-0.7053$$

根据自由度 $n-1=10-1=9$，查 t 分布表得 $t_{\frac{\alpha}{2}}(n-1)=t_{0.025}(9)=2.262$，也可以利用 Excel 中的统计函数 Tiv($\alpha,n-1$)计算得到。

由于 $|t|=0.7053<t_{0.025}(9)=2.262$，所以不拒绝原假设，认为该供货商提供的零件符合要求。

七、总体比例的检验

总体比例是指总体中具有某种相同特征的个体所占的比例，通常用字母 π 表示，用 p 表示样本比例。总体比例的检验与上面介绍的总体均值检验基本上是相同的，区别只在于参数和检验统计量的形式不同。总体均值检验的整个程序都可以作为总体比例检验的参考。这里我们只考虑大样本情形下的总体比例检验。大样本情况下一个总体比例的检验方法如表 8-2 所示。

表 8-2　大样本情况下一个总体比例的检验方法

	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\pi=\pi_0$， $H_1:\pi\neq\pi_0$	$H_0:\pi\geqslant\pi_0$， $H_1:\pi<\pi_0$	$H_0:\pi\leqslant\pi_0$， $H_1:\pi>\pi_0$
检验统计量	$z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$近似服从标准正态分布		
α 与拒绝域	$\|z\|>z_{\frac{\alpha}{2}}$	$z<-z_\alpha$	$z>z_\alpha$
P 值决策准则	$P<\alpha$，拒绝 H_0		

【例 8-14】　一种以休闲和娱乐为主题的杂志，声称其读者群中有 80%为女性。为验证这一说法是否属实，某部门抽取了由 200 人组成的一个随机样本，发现有 146 个女性经常阅读该杂志。分别取显著性水平 $\alpha=0.05$ 和 $\alpha=0.01$，检验该杂志读者群中女性的比例是否为 80%，它们的 P 值各是多少？

解　$H_0: \pi=80\%, H_1: \pi\neq 80\%$。

根据抽样结果计算得 $p=146/200=73\%$，检验统计量为

$$z=\frac{0.73-0.8}{\sqrt{\frac{0.8(1-0.8)}{200}}}=-2.475$$

根据显著性水平 $\alpha=0.05$ 查标准正态分布表得 $z_{\frac{\alpha}{2}}=z_{0.025}=1.96$。由于 $|z|=2.475>z_{\frac{\alpha}{2}}=1.96$，所以拒绝原假设。在显著性水平为 0.05 的条件下，样本提供的证据表明该

杂志的说法并不属实。

根据显著性水平 $\alpha=0.01$ 查标准正态分布表得 $z_{\frac{\alpha}{2}}=z_{0.005}=2.58$。由于 $|z|=2.475<z_{\frac{\alpha}{2}}=2.58$,所以不拒绝原假设。在显著性水平为 0.01 的条件下,样本提供的证据表明该杂志的说法属实。

从该例子可以看出,对于同一个检验,不同的显著性水平将会得出不同的结论,这也是自然的。

八、总体方差的检验

总体方差的检验使用的是卡方(χ^2)分布。此外,总体方差的检验,不论样本容量 n 是大是小,都要求总体服从正态分布,这是由检验统计量的抽样分布决定的。

用 σ_0^2 表示假定的总体方差的某一取值,总体方差假设检验的三种基本形式如表 8-3 所示。

表 8-3 一个总体方差检验的方法

	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\sigma^2=\sigma_0^2$, $H_1:\sigma^2\neq\sigma_0^2$	$H_0:\sigma^2\geqslant\sigma_0^2$, $H_1:\sigma^2<\sigma_0^2$	$H_0:\sigma^2\leqslant\sigma_0^2$, $H_1:\sigma^2>\sigma_0^2$
检验统计量	$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$		
α 与拒绝域	$\chi^2>\chi^2_{\frac{\alpha}{2}}(n-1)$ 或 $\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1)$	$\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1)$	$\chi^2>\chi^2_{\frac{\alpha}{2}}(n-1)$

【例 8-15】 啤酒生产企业采用自动生产线生产罐装啤酒,每瓶的装填量为 640mL,但由于受某些不可控因素的影响,每瓶的装填量会有差异。此时,不仅每瓶的平均装填量很重要,装填量的方差 σ^2 同样很重要。如果 σ^2 很大,会出现装填量太多或太少的情况,这样要么生产企业不划算,要么消费者不满意。假定生产标准规定每瓶装填量的标准差不应超过和不应低于 4mL。企业质检部门抽取了 10 瓶啤酒进行检验,得到的样本标准差为$s=3.8$mL。试以 0.10 的显著性水平检验装填量的标准差是否符合要求。

解 $H_0:\sigma^2=4^2$,$H_1:\sigma^2\neq4^2$

计算检验统计量为

$$\chi^2=\frac{(10-1)\times3.8^2}{4^2}=8.122\,5$$

根据显著性水平 $\alpha=0.10$ 和自由度$(10-1)=9$,查 χ^2 分布表得

$$\chi^2_{0.10/2}(n-1)=\chi^2_{0.05}(10-1)=16.919\,0$$

$$\chi^2_{1-0.10/2}(n-1)=\chi^2_{0.95}(9-1)=3.325\,11$$

由于 $\chi^2_{0.95}(9)=3.325\,11<\chi^2=8.122\,5<\chi^2_{0.05}(9)=16.919\,0$,所以不拒绝原假设 H_0。样本提供的证据表明装填量的标准差符合要求。

第三节 方差分析

一、方差分析有关概念

(一) 方差分析

检验多个总体均值是否相等的统计方法,称为方差分析。

方差分析研究的是分类型自变量对数值型因变量的影响,包括它们之间有没有关系、关系的强度如何等,所采用的方法就是通过检验各总体的均值是否相等来判断分类型自变量对数值型因变量是否有影响。为更好理解,见例 8-16。

【例 8-16】 消费者与产品生产者、销售者或服务的提供者之间经常发生纠纷。当纠纷发生后,消费者常常到消协投诉。为了对几个行业的服务质量进行评价,消费者协会在零售业、旅游业、航空公司、家电制造业分别抽取了不同企业作为样本。其中零售业抽取 7 家,旅游业抽取六家,航空公司抽取 5 家,家电制造业抽取 5 家。每个行业中抽取的这些企业,在服务对象、服务内容、企业规模等方面基本上是相同的。然后统计出最近一年中消费者对总共 23 家企业投诉的次数,结果如图 8-9 所示。

	A	B	C	D	E
1	观测值	行业			
2		零售业	旅游业	航空公司	家电制造业
3	1	57	68	31	44
4	2	66	39	49	51
5	3	49	29	21	65
6	4	40	45	34	77
7	5	34	56	40	58
8	6	53	51		
9	7	44			

图 8-9 23 家企业投诉的次数

被投诉次数越多,说明服务的质量也就越差。消费者协会想知道这几个行业之间的服务质量是否有显著差异。这个问题就是判断"行业"对"投诉次数"是否有显著影响,归结为检验四个行业被投诉的均值是否相等。这就要用到方差分析。

(二) 因素

在方差分析中,所要检验的对象称为因素或因子。例如,在例 8-16 中,"行业"是所要检验的对象,我们就把它称为"因素"或"因子"。

(三) 水平

因素的不同表现称为水平或处理。例如,在例 8-16 中,零售业、旅游业、航空公司、家电制造业是"行业"这一因素的具体表现,我们称之为"水平"或"处理"。

(四) 观测值

每个因子水平下得到的样本数据称为观测值。例如,在例 8-16 中,在每个行业下得到的被投诉次数就是样本数据,也就是观测值。

(五) 随机误差和系统误差

以例 8-16 为例,在同一行业内部,即同一总体中,样本的观察值是不同的。比如在零售业中,所抽取的 7 家企业,被投诉的次数就不完全相同,这是由于在方差分析时,样本是

随机抽取的,因此样本数据之间的差异可以看成是由随机因素的影响造成的,或者说是由于抽样的随机性造成的,我们称之为随机误差。

仍以例 8-16 为例,在不同行业即不同总体下,各观测值也是不同的。这种差异可能源于抽样的随机性,也可能是由于行业本身造成的,后者所形成的误差是由系统性因素造成的,我们称之为系统误差。

(六) 组内误差和组间误差

衡量因素的同一水平(即同一总体)下样本数据的误差,称为组内误差。例如,七家零售业企业被投诉次数之间的误差。

衡量因素的不同水平(不同总体)下各样本之间的误差,称为组间误差。例如,零售业、旅游业、航空公司、家电制造业四个行业之间被投诉次数之间的误差。

组内误差只包含随机误差,组间误差包含随机误差和系统误差。

二、方差分析中的基本假定

方差分析中有以下三个基本假定。

(1) 每个总体服从正态分布。

(2) 各个总体的方差 σ^2 必须相等。

(3) 观测值独立。

三、问题的一般提法

设因素有 k 个水平,每个水平的均值分别用 $\mu_1,\mu_2,\cdots,\mu_k$ 表示,要检验 k 个水平(总体)的均值是否相等,需要提出如下假设。

$H_0: \mu_1=\mu_2=\cdots=\mu_k$,即自变量对因变量没有显著差异和影响;

$H_1: \mu_1,\mu_2,\cdots,\mu_k$ 不全相等,即自变量对因变量有显著差异和影响。

例如,在例 8-16 中,设四个行业被投诉的次数的均值分别是 μ_1,μ_2,μ_3,μ_4。为检验行业对投诉次数有无显著影响,需要提出如下假设。

$H_0: \mu_1=\mu_2=\mu_3=\mu_4$,行业对投诉次数没有显著差异;

$H_1: \mu_1,\mu_2,\mu_3,\mu_4$ 不全相等,行业对投诉次数有显著差异。

从这个例子中可以看到,实际上本节的方法和假设检验一样,但这个问题如果采用两两的假设检验,将要反复进行六次。

四、方差分析步骤

(一) 提出假设

在方差分析中,原假设所描述的是:在按照自变量的值分成的类中,因变量的均值是否相等。因此,检验因素的 k 个水平(总体)的均值是否相等,需要提出如下形式的假设。

$H_0: \mu_1=\mu_2=\cdots=\mu_i=\cdots=\mu_k$,即自变量对因变量没有显著影响;

$H_1: \mu_1,\mu_2,\cdots,\mu_i,\cdots,\mu_k$ 不全相等,即自变量对因变量有显著影响。

如果原假设成立，则自变量对因变量没有显著影响，也就是说自变量与因变量之间没有关系；如果原假设不成立，则自变量对因变量有显著影响，也就是说自变量与因变量之间有关系。

注意：拒绝原假设时，只表明至少有两个总体的均值不相等，并不意味着所有的总体均值都不相等。

（二）构造检验统计量

1. 计算因素各水平（总体）的均值

$$\bar{x}_i = \frac{\sum_{j=1}^{n_j} x_{ij}}{n_i} \quad (i = 1,2,\cdots,k)$$

式中，n_i 为第 i 个总体样本观测值的个数；x_{ij} 为第 i 个总体的第 j 个观测值。

根据图 8-9 中的数据计算零售业的样本均值为

$$\bar{x}_1 = \frac{\sum_{j=1}^{5} x_{ij}}{n_1} = \frac{57+66+49+40+34+53+44}{7}$$

$=49$

使用 Excel 计算均值，可以使用粘贴函数→统计→AVERAGE。

同理，其他行业计算结果如图 8-10 所示。

	A	B	C	D	E
1	观测值	行业			
2		零售业	旅游业	航空公司	家电制造业
3	1	57	68	31	44
4	2	66	39	49	51
5	3	49	29	21	65
6	4	40	45	34	77
7	5	34	56	40	58
8	6	53	51		
9	7	44			
10	样本均值	49	48	35	59
11	样本容量（n_i）	7	6	5	5
12	总均值	$\bar{\bar{x}}=\frac{57+66+\cdots+77+58}{23}=47.869\,565$			
13					

图 8-10 样本均值计算

2. 计算全部观测值的总均值

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$$，其中，$n=n_1+n_2\cdots+n_k$。计算结果如图 8-10 所示。

3. 计算误差平方和，构造检验统计量

为构造检验统计量，共需要计算三个误差平方和：总误差平方和、水平项误差平方和、误差项平方和。

（1）总误差平方和，简记为 SST。它是全部观测值 x_{ij} 与总均值 $\bar{\bar{x}}$ 之间误差的平方和，反映全部观测值的离散状况。

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2$$

本例中，$\bar{\bar{x}}=47.869\,565$，则

$$\text{SST} = (57-47.869\,565)^2 + \cdots + (58-47.869\,565)^2 = 4\,164.608\,696$$

（2）水平项误差平方和，简记为 SSA。它是各组平均值 $\bar{x}_i$ 与总均值 $\bar{\bar{x}}$ 之间误差的平方和，反映各总体的样本均值之间的差异程度，又叫组间平方和。

$$\text{SSA} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{\bar{x}})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

本例计算得

$$SSA = \sum_{i=1}^{4} n_i(\bar{x} - \bar{\bar{x}})^2 = 7 \times (49 - 47.869\,565)^2 + 6 \times (48 - 47.869\,565)^2 + 5 \times (35 - 47.869\,565)^2 + 5 \times (59 - 47.869\,565)^2 = 1\,456.608\,696$$

(3) 误差项平方和,简记为 SSE。它是每个水平或组的各样本数据与其组平均值之间误差的平方和,反映了每个样本各观测值的离散状况,又叫组内平方和或残差平方和。

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$$

本例中,先求出每个行业被投诉的次数与其平均数的误差平方和,然后将四个行业的误差平方和加总,即为 SSE。

零售业:$\sum_{j=1}^{7} (x_{1j} - \bar{x}_1)^2 = (57 - 49)^2 + (66 - 49)^2 + \cdots + (44 - 49)^2 = 700$

旅游业:$\sum_{j=1}^{6} (x_{2j} - \bar{x}_2)^2 = (68 - 48)^2 + (39 - 48)^2 + \cdots + (51 - 48)^2 = 924$

航空公司:$\sum_{j=1}^{5} (x_{3j} - \bar{x}_3)^2 = (31 - 35)^2 + (49 - 35)^2 + \cdots + (40 - 35)^2 = 434$

家电制造业:$\sum_{j=1}^{5} (x_{4j} - \bar{x}_4)^2 = (44 - 59)^2 + (51 - 59)^2 + \cdots + (58 - 59)^2 = 650$

然后将其加总,得 SSE=700+924+434+650=2 708

(4) 三者之间的关系:SST=SSA+SSE。

如果原假设成立,则表明没有系统误差,组间平方和 SSA 除以它的自由度后的均方与组内平方和 SSE 除以它的自由度后的均方差异就不会太大;反之,两者之间的差异就会很大。

SST 的自由度:$n-1$,n 为全部观测值的个数。

SSA 的自由度:$k-1$,k 为因素水平(总体)的个数。

SSE 的自由度:$n-k$。

所以三者的均方分别如下。

SSA 的均方:$MSA=\frac{SSA}{k-1}$,SSE 的均方:$MSE=\frac{SSE}{n-k}$。

所以构造的检验统计量就是

$$F = \frac{MSA}{MSE} \sim F(k-1, n-k)$$

对于本例,$MSA=\frac{SSA}{k-1}=\frac{1\,456.608\,696}{4-1}=485.536\,232$

$$MSE = \frac{SSE}{n-k} = \frac{2\,708}{23-4} = 142.526\,316$$

$$F = \frac{MSA}{MSE} = \frac{485.536\,232}{142.526\,316} = 3.406\,643$$

(三) 统计决策

若 $F>F_\alpha$,则拒绝原假设 H_0,即 $\mu_1=\mu_2=\cdots=\mu_i=\cdots=\mu_k$ 的假设不成立,表明类型自变

量对因变量有显著影响；$F<F_\alpha$，则不能拒绝原假设 H_0，不能认为类型自变量对因变量有显著影响。对于本例，$F=3.406\,643$，假定取显著性水平 $\alpha=0.05$，查表得 $F_{0.05}(3,19)=3.13$，由于 $F>F_\alpha$，所以拒绝原假设，可以认为行业对投诉次数有显著影响。

第四节　样本容量的确定

在抽取样本时样本容量应多大是一个很实际的问题。样本容量取得比较大，收集的信息就比较多，从而估计精度就比较高，但进行观测所投入的费用、人力及时间就比较多；样本容量取得比较小，则投入的费用、人力及时间就比较少，但收集的信息也会比较少，从而估计精度比较低。这说明精度和费用对样本量的影响是矛盾的，不存在既使精度最高又使费用最省的样本量。一个常用的准则是在使精度得到保证的前提下寻求使费用最省的样本量。由于费用通常是样本量的正向线性函数，故使费用最省的样本量也就是使精度得到保证的最小样本量。

一、估计总体均值时样本容量的确定

在简单随机重复抽样下，设样本$(X_1,X_2,\cdots,X_n)$来自正态总体 $N(\mu,\sigma^2)$，总体均值 μ 的点估计为样本均值 $\overline{X}$。如果要求以 $\overline{X}$ 估计 μ 时的绝对误差为 d，可靠度为$(1-\alpha)$，即要求

$$P\{|\overline{X}-\mu|\leqslant d\}=1-\alpha$$

由

$$P\left\{\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right|\leqslant Z_{\frac{\alpha}{2}}\right\}=1-\alpha$$

知

$$P\left\{|\overline{X}-\mu|\leqslant Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\}=1-\alpha$$

故只需取绝对误差

$$d=Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$$

从而解得

$$n=\frac{Z_{\frac{\alpha}{2}}{}^2\cdot\sigma^2}{d^2}\quad（重复抽样条件下）\tag{8-12}$$

同理，在简单随机不重复抽样条件下，我们可以得出估计总体均值时样本容量的计算公式为

$$n=\frac{N\cdot Z_{\frac{\alpha}{2}}{}^2\cdot\sigma^2}{(N-1)d^2+Z_{\frac{\alpha}{2}}{}^2\sigma^2}\quad（不重复抽样条件下）\tag{8-13}$$

【例 8-17】 在某企业中采用简单随机抽样调查职工月平均奖金额，设职工月奖金额服从标准差为 10 元的正态分布，要求估计的绝对误差为 3 元，可靠度为 95%，试问应抽多少名职工？

解　已知 $\sigma=10$，$d=3$，$1-\alpha=0.95$，$1-\frac{\alpha}{2}=0.975$

使用 NORMSINV 函数得$\left[\text{注:在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.025}=1.959\ 963\ 985$$

$$n=\frac{Z_{\frac{\alpha}{2}}{}^{2}\cdot\sigma^{2}}{d^{2}}=\frac{1.96^{2}\times10^{2}}{3^{2}}=42.68\approx43$$

即需抽取 43 名职工作为样本进行调查。

二、估计总体比例时样本容量的确定

在简单随机重复抽样条件下,估计总体比例时,我们可以定义绝对误差 d 为

$$d=Z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\pi(1-\pi)}{n}}$$

从而得到样本容量:

$$n=\frac{Z_{\frac{\alpha}{2}}{}^{2}\cdot\pi\cdot(1-\pi)}{d^{2}}\quad(\text{重复抽样条件下})\tag{8-14}$$

同理,在简单随机不重复抽样条件下,我们可以得出估计总体比例时样本容量的计算公式为

$$n=\frac{N\cdot Z_{\frac{\alpha}{2}}{}^{2}\cdot\pi\cdot(1-\pi)}{(N-1)\cdot d^{2}+Z_{\frac{\alpha}{2}}\cdot\pi\cdot(1-\pi)}\quad(\text{不重复抽样条件下})\tag{8-15}$$

【例 8-18】 根据以往的生产统计,某种产品的合格率为 90%,现要求绝对误差为 5%,在置信水平为 95%的置信区间时,应抽取多少个产品作为样本?

解 已知,$\pi=90\%$, $d=5\%$, $1-\alpha=0.95$, $1-\frac{\alpha}{2}=0.975$

使用 NORMSINV 函数得$\left[\text{注:在 probability 栏内输入}\left(1-\frac{\alpha}{2}\right)\text{值}\right]$

$$Z_{0.025}=1.959\ 963\ 985$$

则 $n=\frac{Z_{\frac{\alpha}{2}}{}^{2}\cdot\pi\cdot(1-\pi)}{d^{2}}=\frac{1.96^{2}\times0.9\times(1-0.9)}{0.05^{2}}\approx139$

第五节 Excel 在抽样推断中的应用

一、参数估计的 Excel 实现

(一) 总体均值的区间估计

对于例 8-1,可以直接使用 CONFIDENCE 函数求置信区间,其操作步骤如下。

单击"公式"菜单下的 f_x(插入函数),在弹出的插入函数对话框中,选择统计函数类别,选择 CONFIDENCE 函数。

在弹出的函数参数对话框中,输入有关参数,如图 8-11 所示。

单击"确定"按钮,就得到 $Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}=3.090\ 995\ 164$。



图 8-11 CONFIDENCE 函数

再用 $\overline{X}=39.5$ 加减 $Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}=3.090\,995\,164$ 即可。

对于例 8-2,可以使用 CONFIDENCE 函数,先求出 $Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$,再乘以 $\sqrt{\frac{N-n}{N-1}}$,最后再与 $\overline{X}$ 加减得到置信区间。

对于例 8-3,可以使用 CONFIDENCE 函数,先求出 $Z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}$,再与 $\overline{X}$ 加减得到置信区间。

(二) 总体比例的区间估计

对于例 8-5,可以选定单元格,输入计算公式直接计算上下限。其操作方法如下。

(1) 选择一个单元格(如 A1),然后输入计算公式"=65%-NORMSINV(0.975)*SQRT(65%*35%/100)",回车后就得到置信区间下限 55.65%。

(2) 选择一个单元格(如 B1),输入计算公式"=65%+NORMSINV(0.975)*SQRT(65%*35%/100)",回车后就得到置信区间上限 74.35%。

对于例 8-6,当然,也可以选定单元格,输入计算公式直接计算上下限。方法同例 8-5。

(三) 总体方差的区间估计

对于例 8-7,可以选定单元格,输入计算公式直接计算上下限。其操作方法如下。

(1) 选择一个单元格(如 A1),然后输入计算公式"=14*1.65/ CHIINV(0.05,14)",回车后就得到置信区间下限 1.61。

(2) 选择一个单元格(如 B1),输入计算公式"=14*1.65/ CHIINV(0.95,14)",回车后就得到置信区间上限 5.8。

二、假设检验的 Excel 实现

(一) 总体均值的假设检验

对于例 8-11,可以利用 P 值进行。P 值可以利用 Excel 中的统计函数功能计算,具体操作步骤如下。

(1) 进入 Excel 表格界面,直接单击“公式”菜单下的 f_x(插入函数)。

(2) 在函数分类中单击“统计”,并在函数名菜单下选择“NORMSDIST”,然后确定。

(3) 将 z 的绝对值 1.01 录入,得到的函数值为 0.843 752 355,该值表示的是在标准正态分布条件下 z 值 1.01 左边的面积。

由于 P 值=0.312 495×[2×(1—0.843 752 355)]远远大于 α=0.05,所以不拒绝 H_0。

对于例 8-12,可以利用 P 值进行。步骤与例 8-11 的步骤完全相同。若直接输入-2.606 1,即可得到 P 值为 0.004 579(因为该命令给出的是分布的左侧面积,恰好就是P 值;若输入 2.606 1,给出的是右侧面积为 0.995 421,P 值则为 1-0.995 421=0.004 579)。

对于例 8-13,t 检验的 P 值同样可以利用 Excel 计算,具体操作步骤如下。

(1) 进入 Excel 表格界面,直接单击“公式”菜单下的 f_x(插入函数)。

(2) 在函数分类中单击“统计”,并在函数名菜单下选择“TDIST”,然后确定。

(3) 在出现对话框的 X 设置中输入计算出的 t 的绝对值 0.705 3。在 Deg-freedom(自由度)设置框中,输入自由度,本例为 9。Tail 设置框中,双侧检验输入 2,单侧检验输入 1,本例输入 2。单击“确定”按钮,即可得到 P 值为 0.498 469 786。

由于 P 值=0.498 469 786>0.05,所以不拒绝原假设。

(二) 总体比例的假设检验

对于例 8-14,可以利用 P 值进行。步骤与例 8-11 的步骤完全相同。由 Excel 计算出的 P 值为 0.013 328。显著性水平为 0.05 时,$P<\alpha=0.05$,拒绝 H_0;显著性水平为 0.01 时,$P>\alpha=0.01$,不拒绝 H_0。结论与统计量检验一致。

(三) 总体方差的假设检验

对于例 8-15,可以利用 P 值进行。若要计算 P 值,可使用 Excel 统计函数中的“CHIDIST”函数。

三、单因素方差分析的 Excel 实现

Excel 操作步骤如下。

(1) 选择“数据”菜单。

(2) 单击“数据分析”功能按钮。

(3) 在分析工具中选择“方差分析:单因素方差分析”选项,再单击“确定”按钮。

（4）当对话框出现时，在“输入区域”设置框内输入数据单元格区域（只输入观测值，本例是从 B3：E9）；在 α 设置框中输入“0.05”（也可根据需要输入其他的显著性水平值）；在“输出选项”中选择输出区域，如图 8-12 所示。



图 8-12　“方差分析：单因素方差分析”对话框

（5）单击“确定”按钮后，得到的输出结果如图 8-13 所示。

	A	B	C	D	E	F	G
1	方差分析：单因素方差分析						
2							
3	SUMMARY						
4	组	观测数	求和	平均	方差		
5	列 1	7	343	49	116.6667		
6	列 2	6	288	48	184.8		
7	列 3	5	175	35	108.5		
8	列 4	5	295	59	162.5		
9							
10							
11	方差分析						
12	差异源	SS	df	MS	F	P-value	F crit
13	组间	1456.609	3	485.5362	3.406643	0.038765	3.12735
14	组内	2708	19	142.5263			
15							
16	总计	4164.609	22				

图 8-13　单因素方差分析结果

图 8-13 中的“方差分析”部分就是方差分析表：“SS”表示平方和；“df”表示自由度；“MS”表示均方；“F”为检验统计量；“P-value”为用于检验的 P 值；“F crit”为给定的 α 水平下的临界值。

从方差分析表中可以看到，由于 $F>F_\alpha$，所以拒绝原假设 H_0，即认为 $\mu_1=\mu_2=\cdots=\mu_i=\cdots=\mu_k$ 不成立，也就是说行业差别对于投诉次数的均值影响是显著的。

在决策时，还可以使用 P 值。若 $P>\alpha$，则不拒绝原假设；若 $P<\alpha$，则拒绝原假设。

思　考　题

1. 简述影响样本容量的因素。
2. 简述假设检验与参数估计的区别与联系。

3. 简述假设检验的步骤。
4. P 值检验和统计量检验的区别有哪些?
5. 方差分析的基本原理有哪些?
6. 单因素方差分析中 SST、SSA 和 SSE 的含义及其关系是什么?
7. 根据方差分析表说明方差分析的步骤。

练 习 题

1. 一家保险公司收集到由 36 个投保人组成的随机样本,得到每个投保人的年龄数据如表 8-4 所示。试建立投保人年龄 90%的置信区间。

表 8-4 某保险公司 36 个投保人的年龄数据 计量单位:岁

23	35	39	27	36	44
36	42	46	43	31	33
42	53	45	54	47	24
34	28	39	36	44	40
39	49	38	34	48	50
34	39	45	48	45	32

2. 一家食品生产企业以生产袋装食品为主,为对食品质量进行监测,企业质检部门经常要进行抽检,以分析每袋重量是否符合要求。现从某天生产的一批食品中随机抽取了 25 袋,测得每袋重量如表 8-5 所示。已知产品重量的分布服从正态分布,且总体标准差为 10 克。试估计该批产品平均重量的置信区间,置信水平为 95%。

表 8-5 某食品企业某天随机抽取 25 袋食品的重量 计量单位:克

112.5	101.0	103.0	102.0	100.5
102.6	107.5	95.0	108.8	115.6
100.0	123.5	102.0	101.6	102.2
116.6	95.4	97.8	108.6	105.0
136.8	102.8	101.5	98.4	93.3

3. 某企业共有职工 1 000 人。企业准备实行一项改革,在职工中征求意见,采取不重复抽样方法随机抽取 200 人作为样本,调查结果显示,有 150 人表示赞成该项改革,有 50 人表示反对。试以 95%的概率确定赞成改革的人数比例的置信区间。

4. 拥有工商管理学士学位的大学毕业生年薪的标准差大约为 2 000 元,假定想要估计年薪 95%的置信区间,希望边际误差为 400 元,应抽取多大的样本量?

5. 根据以往的生产统计,某种产品的合格率约为 90%,现要求边际误差为 5%,在求 95%的置信区间时,应抽取多少个产品作为样本?

6. 一所中学的教务处想要估计试验班和普通班英语考试成绩平均分数差值的置信区间。要求置信水平为 95%,预先估计两个班考试分数的方差分别为:试验班 $\sigma_1^2=90$,普通班 $\sigma_2^2=120$。如果要求估计的误差范围(边际误差)不超过 5 分,在两个班应分别抽取多

少名学生进行调查？

7. 某乐器厂以往生产的乐器采用的是一种镍合金弦线，这种弦线的平均抗拉强度不超过 1 035 MPa，现产品开发小组研究了一种新型弦线，他们认为其抗拉强度得到了提高并想寻找证据予以支持。在对研究小组开发的产品进行检验时，应该采取以下哪种形式的假设？为什么？

（1）$H_0:\mu\leqslant 1\,035$，$H_1:\mu>1\,035$。

（2）$H_0:\mu\geqslant 1\,035$，$H_1:\mu<1\,035$。

（3）$H_0:\mu=1\,035$，$H_1:\mu\neq 1\,035$。

8. 一项调查显示，每天每个家庭看电视的平均时间为 7.25 小时，假定该调查中包括了 200 个家庭，且样本标准差为平均每天 2.5 小时。据报道，10 年前每天每个家庭看电视的平均时间是 6.70 小时，取显著性水平 $\alpha=0.01$，这个调查是否提供了证据支持你认为“如今每个家庭每天收看电视的平均时间增加了”？

9. 经验表明，一个矩形的宽与长之比等于 0.618 的时候会给人们比较良好的感觉。某工艺品工厂生产的矩形工艺品框架的宽与长要求也按这一比例设计，假定其总体服从正态分布，现随机抽取了 20 个框架测得比值数据如图 8-14 所示。

	A	B	C	D
1	0.699	0.615	0.606	0.576
2	0.672	0.611	0.844	0.612
3	0.668	0.57	0.67	0.628
4	0.553	0.654	0.69	0.601
5	0.749	0.606	0.609	0.933

图 8-14　20 个框架测得比值数据

在显著性水平 $\alpha=0.05$ 时能否认为该厂生产的工艺品框架宽与长的平均比例为 0.618？

10. 为比较新旧两种肥料对产量的影响，以便决定是否采用新肥料。研究者选择了面积相等、土壤等条件相同的 40 块田地，分别施用新旧两种肥料，得到的产量数据如图 8-15 所示。

	A	B	C	D	E	F	G	H
1	旧肥料	新肥料	旧肥料	新肥料	旧肥料	新肥料	旧肥料	新肥料
2	109	105	108	99	98	111	104	110
3	98	113	102	103	88	117	100	109
4	103	106	98	118	105	111	104	112
5	97	110	99	99	97	110	106	119
6	101	109	102	107	94	111	101	119

图 8-15　产量数据

显著性水平 $\alpha=0.05$，用 Excel 检验。

（1）新肥料获得的平均产量是否显著地高于旧肥料？假定条件为

① 两种肥料产量的方差未知但相等，即 $\sigma_1^2=\sigma_2^2$。

② 两种肥料产量的方差未知且不相等，即 $\sigma_1^2\neq\sigma_2^2$。

（2）两种肥料产量的方差是否有显著差异？

11. 某家电制造公司准备购进一批 5# 电池，现有 A、B、C 三个电池生产企业愿意供货，为比较它们生产的电池质量，从每个企业各随机抽取五节电池，经试验得其寿命（小时）数据如图 8-16 所示。

试分析三个企业生产的电池的平均寿命之间有无显著差异？如果有差异，用 LSD 方法检验哪些企业之间有差异？（$\alpha=0.05$）

12. 为研究食品的包装和销售地区对其销售量是否有影响，在某市的三个不同地区

中用三种不同包装方法进行销售,获得的销售量数据如图 8-17 所示。

检验不同的地区和不同的包装方法对该食品的销售量是否有显著影响。($\alpha=0.05$)

	A	B	C	D
1	试验号	电池生产企业		
2		A	B	C
3	1	50	32	45
4	2	50	28	42
5	3	43	30	38
6	4	40	34	48
7	5	39	26	40

图 8-16 电池寿命

	A	B	C	D
1	销售地区(A)	包装方法(B)		
2		B_1	B_2	B_3
3	A_1	45	75	30
4	A_2	50	50	40
5	A_3	35	65	50

图 8-17 不同地区不同包装方法的销售量

13. 为检验广告媒体和广告方案对产品销售量的影响,一家营销公司做了一项试验,考察三种广告方案和两种广告媒体,获得的销售量数据如图 8-18 所示。

	A	B	C	D
1			广告媒体	
2			报纸	电视
3	广告方案	A	8	12
4			12	8
5		B	22	26
6			14	30
7		C	10	18
8			18	14

图 8-18 获得的销售量数据

检验广告方案、广告媒体或其交互作用对销售量的影响是否显著。($\alpha=0.05$)

第九章

相关及回归分析

【统计知识 ABC】

[A]"回归"一词的由来

"回归"一词是由英国著名统计学家 Francis Galton 在 19 世纪末期,对人体遗传特征的实验研究孩子及他们的父母的身高时提出来的。

Galton 发现身材高的父母,他们孩子的身高也高,但是这些孩子平均身高却矮于他们的父母;对于身材比较矮的父母,他们的孩子的身高也比较矮,但是这些孩子的平均身高要高于他们的父母。Galton 把这种孩子的身高向中间值靠近的趋势称为一种回归效应,由他发展的研究两个数值变量的方法称为回归分析。

[B]恩格尔系数

恩格尔系数(Engel's Coefficient)是食品支出总额占个人消费支出总额的比重。19 世纪德国统计学家恩格尔根据统计资料,对消费结构的变化得出一个规律:一个家庭收入越少,家庭收入中(或总支出中)用来购买食物的支出所占的比例就越大,随着家庭收入的增加,家庭收入中(或总支出中)用来购买食物的支出比例则会下降。推而广之,一个国家越穷,每个国民的平均收入中(或平均支出中)用于购买食物的支出所占比例就越大,随着国家的富裕,这个比例将呈下降趋势。

联合国根据恩格尔系数的大小,对世界各国的生活水平有了一个划分标准,即一个国家平均家庭恩格尔系数大于 60%为贫穷;50%~60%为温饱;40%~50%为小康;30%~40%属于相对富裕;20%~30%为富足;20%以下为极其富裕。按此划分标准,20 世纪 90 年代,恩格尔系数在 20%以下的只有美国,达到 16%;欧洲、日本、加拿大,一般在 20%~30%,是富裕状态。东欧国家,一般在 30%~40%,相对富裕,剩下的发展中国家,基本上都分布在小康。

改革开放以来,我国城镇和农村居民家庭恩格尔系数已由 1978 年的 57.5%和 67.7%分别下降到 2010 年的 35.7%和 41.1%。

(资料来源:http://baike.baidu.com/view/28093.html.)

[C]外贸依存度

外贸依存度,亦称"外贸依存率"或"外贸系数"。一国对贸易的依赖程度,是开放度的评估与衡量指标。一般用对外贸易额进出口总值在国民生产总值或国内生产总值中所占

比重来表示。即外贸依存度＝对外贸易总额/国民(内)生产总值，或外贸依存度＝净出口额(出口减进口)/国民(内)生产总值。外贸依存度分为出口依存度和进口依存度。出口依存度＝出口总额/国民(内)生产总值，进口依存度＝进口总额/国民(内)生产总值。

2001 年加入世贸组织后，中国外贸依存度逐年上升，2006 年攀至 67%的高点，此后基本呈回落态势。延续 2007 年以来的基本回落趋势，2012 年中国外贸依存度重回 50%以下，为 47%，较上年下降 3.1 个百分点；其中出口依存度为 24.9%，进口依存度为 22.1%，较上年均有回落。但相比美国、日本仍属较高水平。

(资料来源：http://baike.baidu.com/view/694878.htm 和 http://economy.caixin.com/2013-02-07/100490638.html.)

第一节　相关与回归分析的基本问题

一、经济变量间的统计关系及其分类

(一) 经济变量间的统计关系

经济变量之间的数量关系，存在着两种不同的类型，一种是函数关系，另一种是统计关系即相关关系。函数关系是指现象(变量)之间存在着的一种固定的、严格的数量依存关系，即当一个现象(自变量)数值的变动，就会有另一个现象(因变量)完全确定的数值与之对应的变量之间的相互依存关系。函数关系以 $y=f(x)$ 的形式表示，如银行存款中，本利和与本金之间的关系。

相关关系是指现象之间存在着的一种非确定性的数量依存关系，即一个现象发生数量变化时，另一个现象也相应地发生数量变化，但其关系值是不固定的，往往同时出现几个不同的数值，在一定的范围内变动着，这些数值分布在它们的平均数周围的一种数量依存关系。相关关系以 $y=f(x)+u$ 的形式表示，如储蓄额与居民收入之间的关系等。

(二) 相关关系的分类

相关关系的分类有以下几种。

(1) 按相关关系涉及的变量(或因素)的多少，可分为单相关与复相关。单相关也称一元相关，是两个变量之间的相互关系。复相关是指多个变量之间的相互关系，所以复相关又称多元相关。本书主要研究单相关。

(2) 按相关关系的表现形式来分，有线性相关和非线性相关。如果画在直角坐标系上相关的两个变量的对应值，其散布点趋向直线形式，尽管它不是严格的直线关系，但我们还是称其为线性相关或直线相关。例如，施肥量与亩产量之间的关系，在一定的数量界限之内，施肥量增加，亩产量也相应增加，表现为线性相关；但一旦施肥量超过一定的数量，亩产量不但不会增加反而会减少，即出现下降的情况，表现为一种非线性相关。

(3) 按相关的方向，线性相关可分为正相关和负相关。如果两个变量同时趋向在同一方向上变化，即它们是同时增加或同时减少，则称正相关。如经济理论假设商品供给量与商品价格之间具有这种正相关。当价格上涨时，供给量就增加；当价格下跌时，供给量

就减少。反之,如果两个变量不在同一方向上变化,即一个变量增加,而另一个变量减少,呈反向变化,则称为负相关,如商品需求量和商品价格是负相关。当商品价格上涨时,需求量减少,而当商品价格下降时,需求量增加。

(4) 按变量之间的相关程度来分,可分为完全相关、不完全相关和不相关三类。所谓完全相关,就是变量之间的一种确定性的函数关系。反之,若变量之间不存在相关关系,彼此独立,相互之间没有联系,则称为不相关。介于两者之间的称为不完全相关。

二、相关分析的主要内容

相关分析是研究两个或两个以上的变量之间相关程度的大小的一种统计方法。其主要内容如下。

(1) 确定现象之间有无关系存在,以及相关关系呈现的形态。

(2) 确定相关关系的密切程度。判断相关关系密切程度的主要方法是绘制散点图和计算相关系数。

(3) 相关系数的检验。由于两个变量的相关系数大多是由样本值计算出来的,即用两变量的样本相关系数来描述两变量(总体)的相关性。这样就产生了如下问题:样本相关系数的绝对值大到什么程度才能断定 X 与 Y 间可能存在线性关系呢?这就需要进行显著性检验。

三、回归分析的主要内容

回归分析是寻找具有相关关系的变量间的数学表达式并进行统计推断的一种统计方法。其主要内容包括以下几点。

(1) 进行参数估计。即如何根据样本观测值对回归模型的参数进行估计,求出具体的回归方程。

(2) 进行统计显著性检验。即对回归方程、参数估计值进行显著性检验与校正,以便使回归方程或参数更加优良。

(3) 进行预测和控制。如何根据回归方程进行适当的预测和控制是回归分析的最终目的。

四、相关分析与回归分析二者的关系

(一) 二者的联系

相关分析与回归分析都是研究和处理变量之间相关关系的数理统计方法。

相关分析和回归分析是互相补充、密切联系的。回归分析是建立在相关分析的基础上,对于具有密切相关的两个变量进行深入分析,建立它们之间的数学关系式,并进行统计推断,是相关分析的拓展。而相关分析是回归分析的前提,对于相关程度很低的两个变量进行回归分析是没有实际意义的。

(二) 二者的区别

相关分析与回归分析的区别有以下几点。

(1) 相关分析主要通过相关系数来判断两个变量之间是否存在着相互关系及其关系的密切程度,其前提条件是两个变量都是随机变量,且变量之间不必区别自变量和因变量,而回归分析研究一个随机变量(Y)与另一个非随机变量(X)之间的相互关系,且变量之间必须区别自变量和因变量。

(2) 相关系数只能观察变量间相关关系的密切程度和方向,不能估计推算具体数值。而回归分析可以根据回归方程,用自变量数值推算因变量的估计值。

(3) 互为因果关系的两个变量,可以拟合两个回归方程,且互相独立、不能互相替换。而相关系数却只有一个,即自变量与因变量互换相关系数不变。

第二节 相关分析

一、定性分析

从数量上研究社会经济现象的依存关系,首先要凭借研究者所掌握的科学知识、判断能力进行定性分析。定性分析需根据马克思主义哲学关于事物普通联系和相互作用的原理和社会经济理论进行分析研究,否则就很有可能将虚假的相关现象拿来进行相关、回归分析,其内容则会变成抽象的数字游戏,其结果将导致预测和决策失误。

二、相关图表

在定性分析的基础上,对样本资料编制相关表和绘制相关图,可以直观地判断现象之间大致上呈现何种关系的形式,粗略地研究变量间是否存在着相关关系以及相关关系的方向和密切程度。

(一) 相关表

相关表是根据现象变动样本资料编制出来的反映变量间相关关系的统计表。根据样本资料是否分组,相关表可分为简单相关表和分组相关表。

(1) 简单相关表是资料未曾分组,只将自变量的取值按照从小到大的顺序并配合因变量的取值一一对应平等排列起来的表。其编制程序是:将相关资料中的两个变量,分为自变量和因变量,将两个变量值一一对应,按照自变量的值从小到大顺序排列即成。

(2) 分组相关表是指将原始资料进行分组而形成的相关表。可分为单变量分组相关表和双变量分组相关表。对于单变量分组相关表,对自变量进行分组并计算次数,而对因变量分组,只计算其平均值。根据资料的具体情况,自变量分组可以是单项式,也可以是组距式。单变量分组相关表与简单相关表比较,更能清晰地反映出两变量之间的相关关系。双变量分组相关表是指自变量和因变量都进行分组而形成的相关表,这种表形似棋盘,故又称为棋盘式相关表。

(二) 相关图

利用直角坐标系第一象限,把自变量置于横轴上,因变量置于纵轴上,将两变量相对

应的变量值用坐标点形式描绘出来，用以表明相关关系的图形，称为相关图。相关图的作用有以下几点。

（1）判断现象之间有无相关关系（图 9-1）。

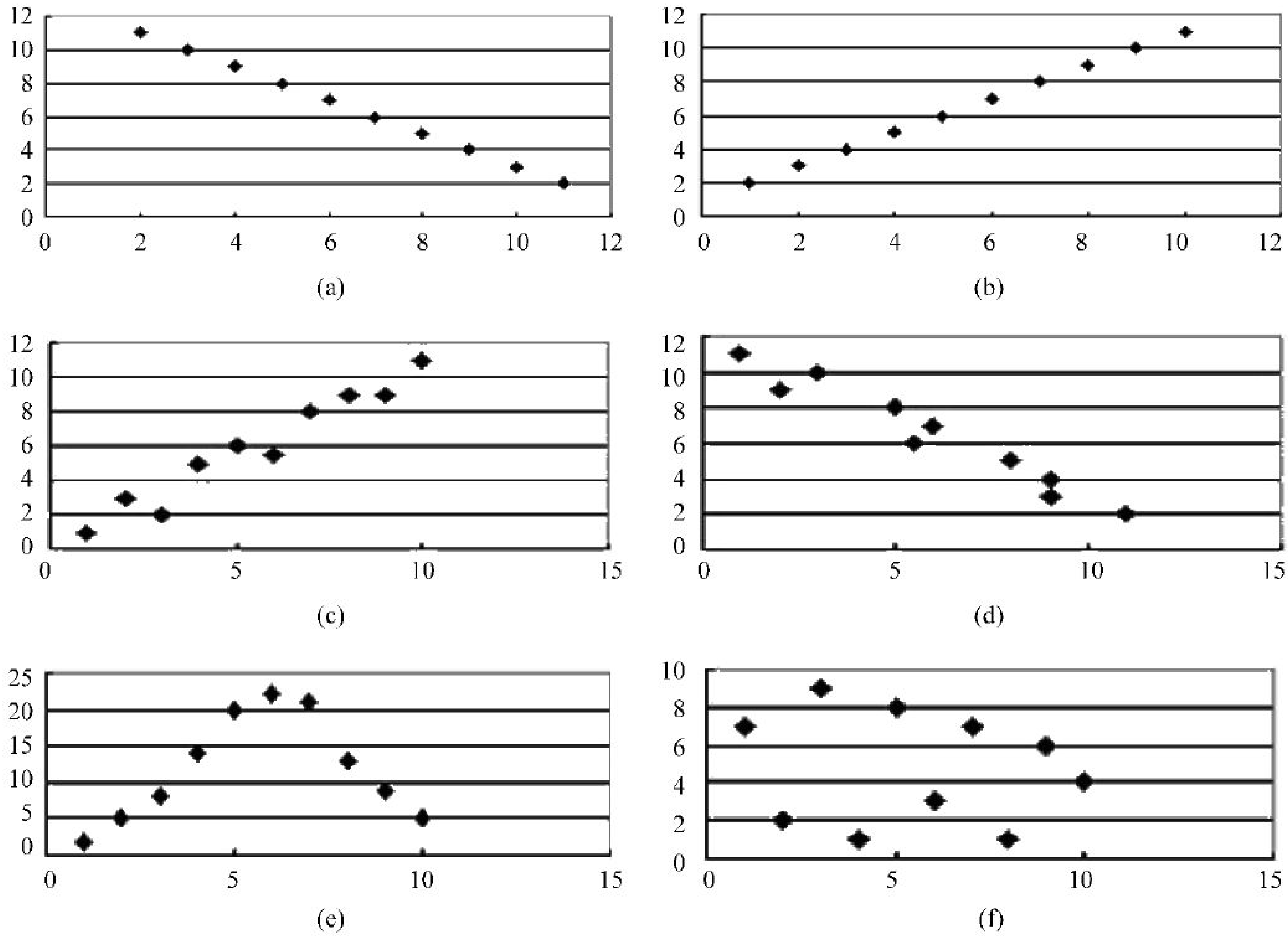


图 9-1　相关关系的类型

（a）完全线性负相关；（b）完全线性正相关；（c）不完全线性正相关；（d）不完全线性负相关；（e）非线性相关；（f）不相关

（2）观察相关关系的类型（图 9-1）。

（3）观察相关关系的密切程度（图 9-2）。

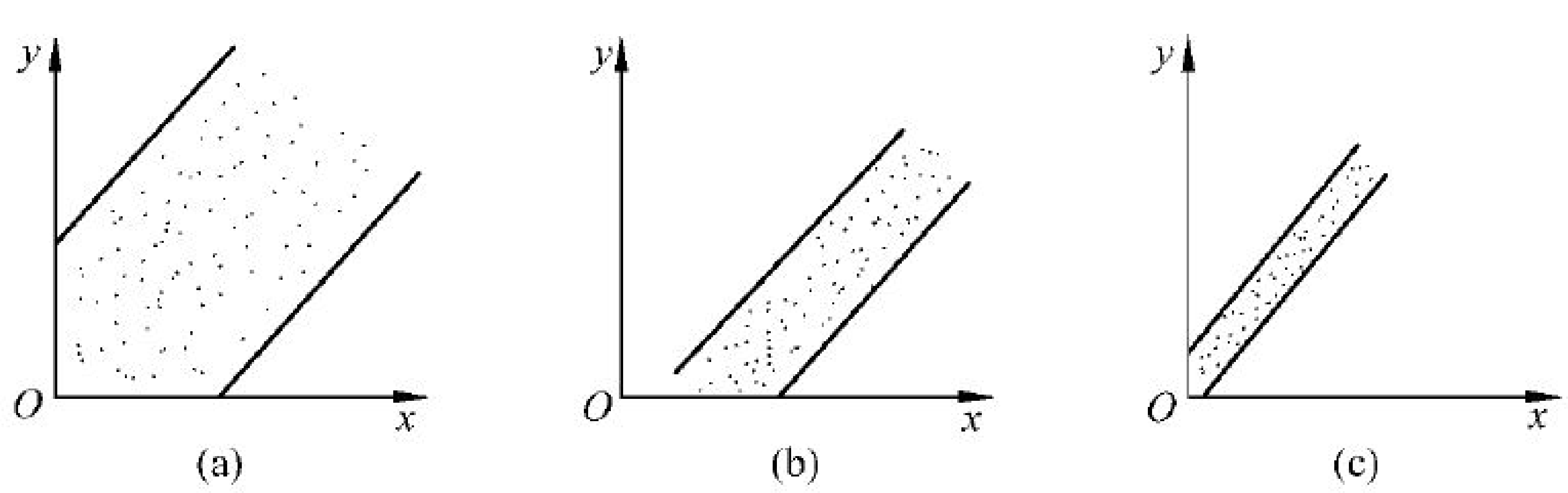


图 9-2　相关关系的密切实程度

（a）关系比较松散；（b）关系比较紧密；（c）关系很紧密

三、相关系数

相关图表只能粗略地大体上反映变量间相关关系的方向、形式和密切程度,要确切地反映相关关系的密切程度,还需计算相关系数。

在各种相关中,单相关是基本的相关关系,它是复相关的基础。单相关有线性和非线性相关两种表现形式。测定线性相关系数的方法是最基本的相关分析,是测定其他相关系数方法的基础。我们着重研究线性的单相关系数即直线相关系数,简称相关系数。

(一) 皮尔逊相关系数

皮尔逊(Pearson)相关系数的计算公式为

$$r=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x^2)}\sqrt{n\sum y^2-(\sum y^2)}}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\sigma_x\sigma_y} \tag{9-1}$$

式中,$\sigma_{xy}=\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$,称为 xy 的协方差;$\sigma_x=\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}$,是变量 x 的标准差;$\sigma_y=\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}$,是变量 y 的标准差。

(二) 相关系数的性质

相关系数的性质有以下几点。

(1) 当 $|r|\leqslant 1$ 时,即相关系数是介于 -1 和 $+1$ 之间的实数。

(2) 当 $|r|=1$ 时,变量 x 与 y 为完全线性相关,即变量 x 与 y 之间存在着确定的函数关系。

(3) 当 $0<|r|<1$ 时,表明变量 x 与 y 之间存在着一定的线性相关关系。$|r|$ 的数值越接近于 1,说明 x 与 y 之间线性相关程度越高;反之 $|r|$ 的数值越接近于 0,说明 x 与 y 之间线性相关程度越低。

(4) 当 $r>0$ 时,表明 x 与 y 为正相关;当 $r<0$ 时,表明 x 与 y 为负相关。

(5) 当 $r=0$ 时,表明 x 与 y 之间没有线性相关关系,即 x 与 y 之间不相关或曲线相关。

(三) 斯皮尔曼等级相关系数

对于分类数据、顺序数据,可以计算斯皮尔曼(Spearman)等级相关系数。等级相关系数是把数据转换成等级,测定等级之间的相关程度。

斯皮尔曼等级相关系数的计算公式为

$$r=1-\frac{6\sum d^2}{n(n^2-1)} \tag{9-2}$$

式中,d 为两变量取值对应等级的等级差;$\sum d^2$ 为两变量所有等级差的平方和;n 为等级项数。

四、相关系数检验

相关系数是根据样本数据计算出来的，两个不相关的变量，其样本相关系数也可能较高，这在统计上称为虚假相关。要从样本相关系数判断总体是否也具有这样的关系，则需要对相关系数进行统计检验后才能得出结论。

相关系数描述 x 与 y 之间的密切程度与样本个数 n 有关，当 n 较小时，相关系数的绝对值容易接近 1，特别地当 $n=2$ 时，相关系数的绝对值一定为 1。因此，我们并不能根据相关系数的绝对值大小直接衡量 x 与 y 之间关系是否真正密切，还必须通过相关系数临界值进行比较。

(1) 提出原假设 $H_0: r=0$，备择假设 $H_1: r\neq 0$。

(2) 构造相关系数统计量 r

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x)^2}\sqrt{n\sum y^2-(\sum y)^2}}$$

(3) 给定一个小概率(显著性水平)α，查相关系数临界值表得 $r_\alpha(n-m)$，$m=(k+1)$(k 为自变量的个数)为变量的个数或估计参数个数，在一元线性回归方程中 $m=2$。

(4) 做出统计决策。若 $|r|\geqslant r_\alpha(n-m)$，则拒绝原假设，接受备择假设，此时我们可认为 x 与 y 之间线性相关关系显著；若 $|r|<r_\alpha(n-m)$，则接受原假设，此时我们可认为 x 与 y 之间的线性相关关系不显著。

第三节　线性回归分析

一、线性回归模型

（一）回归模型的一般形式

对于模型 $y=f(x)+u$，若 x 与 y 确有因果关系，则称 $y=f(x)+u$ 为总体回归模型，x 为自变量(或解释变量或外生变量)，y 为因变量(或被解释变量或内生变量)，u 为随机项，是没有包含在模型中的自变量和其他一些随机因素对 y 的总影响。

若给定 x,y 的 n 次观察值(样本值)(x_i,y_i)，$i=1,2,3,\cdots,n$ 代入模型 $y=f(x)+u$ 中得

$$y_i=f(x_i)+u_i \tag{9-3}$$

我们称式(9-3)为样本回归模型。

一般说来，随机项 u 来自以下几个方面。

(1) 自变量的省略。由于人们认识的局限，不能穷尽所有的影响因素，或由于受时间、费用、数据质量等制约，而没有引入模型之中的，而对被解释变量 y 又有一定影响的自变量被省略。

(2) 统计误差。数据收集中由于计量、计算、记录等导致的登记误差；或由样本信息推断总体信息时产生的代表性误差。

(3) 模型的设定误差。如在模型构造时,非线性关系用线性模型描述了;复杂关系用简单模型描述了;此非线性关系用彼非线性模型描述了;等等。

(4) 随机误差。被解释变量还受一些不可控制的众多的、细小的偶然因素的影响。

(二) 一元线性回归模型

对于总体回归模型 $y=f(x)+u$,特别地,当 $f(x)=\beta_0+\beta_1 x$ 时,则有

$$y=\beta_0+\beta_1 x+u$$

式中,β_0 和 β_1 为两个待定参数;β_0 为直线的截距;β_1 为直线的斜率。我们称之为一元线性总体回归模型。

若给定 x,y 的 n 次观察值(样本值)$(x_i,y_i),i=1,2,3,\cdots,n$ 代入模型 $y=\beta_0+\beta_1 x+u$ 中得

$$y_i=\beta_0+\beta_1 x_i+u_i \quad (i=1,2,3,\cdots,n) \tag{9-4}$$

我们称式(9-4)为一元线性样本回归模型。

二、线性回归模型中随机项的基本假定

在给定样本观测值(样本值)$(x_i,y_i),i=1,2,3,\cdots,n$ 后,为了估计式(9-4)的参数 β_0 和 β_1,必须对随机项 u_i 做出某些合理的假定,这些假定通常称为古典假设。

假设 1 $E(u_i)=0$。即随机项 u_i 的数学期望(均值)为零。

这就是说,对于 x 的每个观测值 u 可以取不同值,有的大于零,有的小于零,但对于 u 的所有可能取值,它们的平均数均等于零。

假设 2 $\mathrm{Var}(u_i)=E((u_i-E(u_i))^2)=E(u_i^2)=\sigma_u^2(i=1,2,\cdots,n)$。即各次观测中 u 具有相同的方差,也就是说各次观测所受的随机影响的程度相同。

对于不同的解释变量 x_i,如果随机项的方差不同,那么与其相对应的观察值 y_i 的可靠程度(与随机项的方差大小成反比)也不相同。对于不同的随机项 u_i 所对应的不同观测值 y_i 应分别赋予不同的权数,这样做会使参数的估计、检验和利用模型进行预测复杂化。如果满足同方差假设,将会使估计、检验和预测简化。

假设 3 $\mathrm{Cov}(u_i,u_j)=E((u_i-E(u_i))(u_j-E(u_j)))=0 \quad (i\neq j;\ i=1,2,\cdots,n;\ j=1,2,\cdots,n)$。即在任意两次观测时,$u_i,u_j$ 是相互独立的、不相关的,也就是无序列相关。

如果这个假设成立,参数的检验和利用模型进行预测将被简化。

假设 2、3 称为高斯-马尔可夫(Gauss-Markov)假设。在此假设条件下,可以得到关于回归系数的最小二乘估计及随机项方差估计的一些重要性质。

假设 4 $\mathrm{Cov}(u_i,x_i)=0$。即解释变量 x_i 与误差项 u_i 同期独立无关。如果两者相关,就不可能把 x 对 y 的影响和 u 对 y 的影响区分开来。

在一般情况下,x_i 为非随机变量(在预测时它是确定性变量),而 u_i 为随机变量,这一假定很显然成立。

假定 5 $u_i\sim N(0,\sigma_u^2)$。即 u_i 为服从正态分布的随机变量。

对于大样本,由中心极限定理中的李雅普诺夫(Liapunov)定理知,无论 u_i 中包含的每一种影响因素服从什么分布,u_i 都近似正态分布,即在大样本条件下这个假设成立。

但对于小样本，这个假设不一定成立，如果这个假设不成立，就无法进行检验和预测。因为检验和预测，须知道总体 y 的分布情况。

在 u_i 为服从正态分布的假定下，随机变量 y_i 也服从正态分布：

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_u^2), \quad i = 1, 2, \cdots, n$$

对于多元线性回归模型还需假定各解释变量之间不存在显著的线性相关关系。换句话说就是自变量之间不存在多重共性性。即 $\mathrm{rank}(X) = k+1 < n$，也就是说矩阵 $\boldsymbol{X}$ 的秩等于参数个数。

三、一元线性样本回归方程

对于一元线性回归模型 $y_i = \beta_0 + \beta_1 x_i + u_i$，在满足古典假设条件下，两边取均值得一元线性理论回归方程 $E(y_i) = \beta_0 + \beta_1 x_i$，简称总体回归线。其中 β_0 和 β_1 是未知的，实际上总体回归线是无法求得的，它只是理论上的存在，所以称为理论回归方程。但我们可以通过样本观测值来拟合一条直线，使它成为理论回归线的最佳估计：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 \tag{9-5}$$

式中，$\hat{\beta}_0$ 为估计的回归直线在 y 轴上的截距；$\hat{\beta}_1$ 为直线的斜率，它表示对于一个给定的 x 值；$\hat{y}$ 为 y 的估计值。即 $\hat{\beta}_1$ 表示 x 每变动一个单位时，y 的平均变动值。

这条线我们称为一元线性样本回归线，简称样本回归线；式(9-5)称为一元线性样本回归方程，简称样本回归方程。又因式(9-5)的建立依赖于观测值，所以我们又称式(9-5)为经验回归方程。

四、普通最小平方法

观测值 y_i 与它的拟合值 $\hat{y}_i$ 之差 e_i 叫作残差，它是随机项 u_i 的估计值。记为

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

最小二乘准则是使全部观测值的残差平方和为最小，即

$$Q = \sum e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \mathrm{Min}$$

由微分极值原理知，要使 Q 达到最小，必要条件是：Q 对 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的一阶偏导数等于零，二阶偏导数大于零。依据必要条件，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 应满足下列方程组：

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \end{cases}$$

经整理后得正规方程组：

$$\begin{cases} \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{cases}$$

解之得

$$\begin{cases} \hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{9-6}$$

式中，$\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$，$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$。

说明回归直线 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 通过平均数这个点($\bar{x}$，$\bar{y}$)。

由式(9-6)得到的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 称为最小二乘估计量(ordinary least square estimators，OLSE)。

上述方法就是普通最小平方法(ordinary least square method)，简记为 OLS。

五、统计显著性检验

当给定 x，y 的 n 次观测值(样本值)(x_i, y_i) $i=1,2,\cdots,n$，采用最小平方法进行参数估计得经验回归方程 $\hat{y}=b_0+b_1 x$ 后，我们还不能马上就用它去进行经济分析、预测和控制。因为 $\hat{y}=b_0+b_1 x$ 是否真正描述了变量 x 与 y 之间的统计关系，或变量 x 与 y 之间的统计关系是否显著，还需对有关参数和方程进行统计显著性检验，才能做出回答或判定。

(一) 回归系数显著性检验——t 检验

t 检验属于回归系数的统计显著性检验，是对个别参数显著性的检验。以斜率系数 β_1 的检验为例，其步骤为

(1) 提出假设。

原假设 H_0：$\beta_1=0$，备择假设 H_1：$\beta_1 \neq 0$。

如果原假设 H_0 成立，则因变量 y 与自变量 x 之间并没有真正的线性相关关系，即无论变量 x 怎样变化，因变量 y 始终等于 β_0，也就是说自变量 x 的变化对因变量 y 并不产生影响。

(2) 构造 t 统计量。

$$t = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} \sim t(n-m) \tag{9-7}$$

式中，$S(\hat{\beta}_1) = \sqrt{\dfrac{\hat{\sigma}_u^2}{\sum(x-\bar{x})^2}}$ 为 $\hat{\beta}_1$ 的标准差；$\hat{\sigma}_u^2 = \dfrac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 为 $\hat{\sigma}_u^2$ 的无偏估计量。

(3) 给定小概率(显著水平 α)，查 t 分布临界值 $t_{\alpha/2}(n-2)$。

(4) 做出统计决策。

当 $|t| \geqslant t_{\alpha/2}(n-2)$，拒绝原假设，认为 β_1 显著不为零，说明因变量 y 对自变量 x 的一

元线性相关关系显著；当$|t|<t_{\alpha/2}(n-2)$，不能拒绝原假设，认为β_1与零没有显著差异，说明因变量y对自变量x的一元线性相关关系不显著。

（二）回归方程的显著性检验——F检验

F检验属于回归方程的显著性检验，它是对所有参数感兴趣的一种显著性检验。对于一元线性回归模型，我们感兴趣的参数只有β_1。其检验步骤为：

（1）提出假设。

原假设$H_0: \beta_1=0$，备择假设$H_1: \beta_1\neq 0$。

（2）构造F统计量。

我们把总离差平方和定义为：$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$；回归解释平方和定义为：$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$；残差平方和定义为：$SST=\sum_{i=1}^{n}(y_i-\hat{y})^2$。则

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE}\sim F(1,n-2) \tag{9-8}$$

即统计量F服从第一自由度为1，第二自由度为$(n-2)$的F分布。

（3）给定显著水平α，查F分布临界值$F_\alpha(1,n-2)$。可以利用Excel中的粘贴函数“Finv”进行计算。

（4）做出统计决策。

若$F\geqslant F_\alpha(1,n-2)$时，拒绝原假设$H_0$，接受备择假设，则认为$x$与$y$的线性相关关系显著即回归方程显著；若$F<F_\alpha(1,n-2)$时，接受$H_0$，则认为$x$与$y$的线性相关关系不显著，即回归方程不显著。

若我们知道了Significance F，通常也称为P值，也可根据P值比较，做出决策。其判别标准是：

当Significance F$\leqslant\alpha$时，$F\geqslant F_\alpha(1,n-2)$，此时拒绝原假设$H_0$；当Significance F$>\alpha$时，$F<F_\alpha(1,n-2)$，此时应接受原假设$H_0$。

（三）变量x与变量y之间的线性相关关系检验——r检验

由于一元线性回归方程研究的是变量x与变量y之间的线性相关关系，所以我们可以用反映变量x与变量y之间的相关关系密切程度的相关系数来检验回归方程的显著性。具体检验步骤见本章第二节。

（四）拟合优度检验——R^2检验

因变量总离差平方和可做如下分解：

$$y_i-\bar{y}=(\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)=(\hat{y}-\bar{y})+e_i$$

对上式两边求平方和得

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

由此可见，总离差平方和（total sum of squares，TSS）可以分为可解释平方和

(explaned sum of squares,ESS)与残差平方和(residualum of squares, RSS)两部分。即

$$\mathrm{TSS}=\mathrm{ESS}+\mathrm{RSS}$$

式中,$\mathrm{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$

$$\mathrm{ESS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_0\sum_{i=1}^{n}y_i+\hat{\beta}_1\sum_{i=1}^{n}x_iy_i-n\bar{y}^2$$

$$\mathrm{RSS}=\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}y_i^2-\hat{\beta}_0\sum_{i=1}^{n}y_i-\hat{\beta}_1\sum_{i=1}^{n}x_iy_i$$

ESS 是由回归方程确定的,也就是由自变量 x 变动引起的,又称为回归平方和;RSS 是由 x 之外的随机项 u 的波动引起的,又称不可解释平方和。不难看出,回归平方和(可解释平方和)ESS 在总平方和 TSS 中所占比例越大,残差平方和 RSS 在 TSS 中所占比重就越小,说明回归的效果就越好,即样本回归线 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$ 与样本观测值(x_i,y_i)拟合得越好。为此我们把回归平方和占总平方和的比重定义为样本决定系数,记为

$$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}=1-\frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{9-9}$$

显然 $0\leqslant R^2\leqslant 1$。$R^2$ 的数值等于相关系数的平方。R^2 越接近于 1,表示回归直线与样本观测值拟合越好。可见 R^2 可以用来度量回归直线与样本观测值拟合优度。另外,若 R^2 大,则解释变量 x 对被解释变量 y 的解释程度就高,可以推测总体线性相关关系显著,即总体回归系数 β_1 不会同时为零,回归方程显著。反之,可以推测总体线性相关关系不显著,即 β_1 与零没有显著差异,回归方程不显著。

样本决定系数与总体回归系数 β_1 有如下关系:

$$R^2=\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}=\hat{\beta}_1^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

(五)四种检验的关系

前面介绍了 t 检验、F 检验、r 检验和 R^2 检验,对于一元线性回归方程来说,这四种检验是等价的。

可以证明:$t=\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$,$r=\sqrt{\dfrac{F}{(n-m)+F}}$,$F=t^2$,$\sqrt{R^2}=r$

因此,对于一元线性回归方程,我们只需做其中的一种检验即可。但对于多元线性回归方程这四种检验有着不同的意义,并不是等价的,需分别进行检验。

六、预测

当参数估计出来后,在古典假设满足的条件下,若通过了各种统计显著性检验,并具有经济含义,便可以用估计出来的回归方程进行预测了。

(一)点预测

点预测分为两种:一是平均值的点预测,二是个别值的点预测。利用回归方程,对于

x 的一个固定值 x_0，推算出 y 的平均值的一个估计值 $E(y_0)$，就是平均值的点预测；如果对于 x 的一个特定值 x_0，推算出 y 的一个个别值的估计值 $\hat{y}_0$，则属于个别值的点预测。

对于给定的 x_0，则 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$，$E(y_0)=\beta_0+\beta_1 x_0$。

（二）区间预测

区间预测是指，对于给定的显著水平 α，找一个区间 (y_1,y_2)，使对应于某个特定的 x_0 的实际值 y_0，以 $(1-\alpha)$ 的概率被区间 (y_1,y_2) 所包含。即

$$(\hat{y}_0-y_1,\hat{y}_0+y_2) \quad 满足 \quad P\{y_1<y_o<y_2\}=1-\alpha$$

对应于点预测，区间预测也分为两种：一种是平均值的区间预测，另一种是个别值的区间预测。

1. 个别值的区间预测

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{9-10}$$

2. 平均值的区间预测

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \tag{9-11}$$

由区间预测公式可以看出，对于给定的显著水平 α，样本容量 n 越大，$\sum_{i=1}^{n}(x_i-\bar{x})^2$ 就越大，x_0 越靠近 $\bar{x}$，则置信区间长度就越短，此时的预测精度就高。所以，为了提高预测精度，样本容量 n 应越大越好，所给定的 x_0 不能偏离 $\bar{x}$ 太大。当 $x_0=\bar{x}$ 时，预测结果精度最高；当 $|x_0-\bar{x}|$ 很大时，预测效果就差。

七、案例：一元线性回归模型的应用

【例 9-1】 某公司 12 个汽车销售分公司的有关数据如表 9-1 所示。试进行一元线性回归分析。

表 9-1 某公司 12 个汽车销售分公司的销售量和广告费资料

分公司名称	汽车销售量 y/辆	广告费 x/万元	x^2	xy	y^2
A	1 000	357	127 449	357 000	1 000 000
B	1 100	385	148 225	423 500	1 210 000
C	1 250	420	176 400	525 000	1 562 500
D	1 280	406	164 836	519 680	1 638 400
E	1 360	490	240 100	666 400	1 849 600
F	1 480	525	275 625	777 000	2 190 400
G	1 500	602	362 404	903 000	2 250 000
H	1 720	651	423 801	1 119 720	2 958 400

续表

分公司名称	汽车销售量 y/辆	广告费 x/万元	x^2	xy	y^2
I	1 800	735	540 225	1 323 000	3 240 000
J	1 890	721	519 841	1 362 690	3 572 100
K	2 100	840	705 600	1 764 000	4 410 000
M	2 200	924	853 776	2 032 800	4 840 000
合计	18 680	7 056	4 538 282	11 773 790	30 721 400

(一)建立模型

由表 9-1 计算知：$n=12$，$\sum_{i=1}^{12}x_i=7\ 056$，$\sum_{i=1}^{12}y_i=18\ 680$，$\sum_{i=1}^{12}x_i^2=4\ 538\ 282$，$\sum_{i=1}^{12}x_iy_i=11\ 773\ 790$，$\sum_{i=1}^{12}y_i^2=30\ 721\ 400$。将其代入式(9-6)得

$$\begin{cases}\hat{\beta}_1=\dfrac{12\times 11\ 773\ 790-7\ 056\times 18\ 680}{12\times 4\ 538\ 282-7\ 056\times 7\ 056}=2.028\ 873\\ \hat{\beta}_0=\dfrac{1}{12}\times 18\ 680-\dfrac{1}{12}\times 7\ 056=363.689\ 1\end{cases}$$

于是得一元线性回归方程为

$$\hat{y}_i=363.689\ 1+2.028\ 873x_i$$

(二)统计显著性检验

1. 对于 t 检验

$$S(\hat{\beta}_1)=\sqrt{\frac{\hat{\sigma}_u^2}{\sum(x_i-\bar{x})^2}}=\sqrt{\frac{4\ 015.81}{389\ 354}}=0.101\ 558$$

$$t=\frac{\hat{\beta}_1}{S(\hat{\beta}_1)}=\frac{2.028\ 873}{0.101\ 558}=19.977\ 487$$

给定显著水平 $\alpha=5\%$，查 t 分布临界值 $t_{0.05/2}(10)=1.812$，因为 $t=19.977\ 487>t_{0.05/2}(10)=1.812$，所以我们拒绝原假设 $H_0:\beta_1=0$，接受备择假设 H_1，认为 β_1 显著不为零，说明因变量 y 对自变量 x 的一元线性相关关系显著，或 β_1 显著。

2. 对于拟合优度检验

$$\text{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2=1\ 642\ 866.7$$

$$\text{ESS}=\sum_{i=1}^{n}(\hat{y}-\bar{y})^2=\hat{\beta}_0\sum_{i=1}^{n}y_i+\hat{\beta}_1\sum_{i=1}^{n}x_iy_i-n\bar{y}^2=1\ 602\ 708.6$$

$$\text{RSS}=\sum_{i=1}^{n}(y_i-\hat{y})^2=\text{TSS}-\text{ESS}=4\ 015.807\ 1$$

$$R^2=\frac{\text{ESS}}{\text{TSS}}=1-\frac{\text{RSS}}{\text{TSS}}=0.987$$

拟合优度较高。

（三）预测

若要估计广告费用为 1 000 万元时，所有 12 个汽车销售分公司的汽车销售量的平均数为 $E(y_0)=363.689\,1+2.028\,873\times 1\,000=2\,393$(辆)，就是平均值的点预测；若要估计广告费用为 602 万元的那个汽车销售分公司的汽车销售量为 $\hat{y}_0=363.689\,1+2.028\,873\times 602=1\,585$(辆)就属于个别值的点预测。

若要估计广告费用为 1 000 万元时，所有 12 个汽车销售分公司的汽车销售量的平均数区间预测为

$$2\,393 \pm t_{0.05/2}(10)\times\sqrt{\frac{1}{12}+\frac{(1\,000-588)^2}{389\,354}}\times\sqrt{4\,015.81}$$

$$2\,393 \pm 1.812\times 0.720\,62\times 63.370\,4$$

即 2 310 ～ 2 476　(以 95% 的概率保证)

这就是平均值的区间预测。

广告费用为 602 万元的那个汽车销售分公司的汽车销售量区间预测为

$$1\,585 \pm t_{0.05/2}(10)\times\sqrt{1+\frac{1}{12}+\frac{(1\,000-588)^2}{389\,354}}\times\sqrt{4\,015.81}$$

$$1\,585 \pm 1.812\times 1.232\,6\times 63.370\,4$$

即 1 443 ～ 1 727　(以 95% 的概率保证)

这就是个别值的区间预测。

第四节　非线性回归分析

一、非线性回归

实际工作中，有时两个变量之间的相关关系并非线性关系，而是呈现某种非线性(曲线)的关系。这时，我们必须考虑根据恰当的曲线类型为两个变量配合一条相应的曲线做回归分析。

在许多情况下，非线性回归问题可以通过变量的变换化成线性回归问题，然后应用前面介绍过的线性回归分析方法来解决非线性回归问题。

【例 9-2】　某地区 2001—2016 年工业总产值和投资额资料如表 9-2 所示。

表 9-2　某地区 1993—2008 年工业总产值和投资额资料　　亿元

年　份	工业总产值	投资额	年　份	工业总产值	投资额
2001	5.23	0.016	2009	9.35	0.031
2002	5.63	0.015	2010	10.63	0.034
2003	5.94	0.016	2011	10.63	0.034
2004	6.35	0.019	2012	11.71	0.035
2005	6.88	0.025	2013	13.06	0.044
2006	7.53	0.029	2014	14.13	0.056
2007	7.96	0.028	2015	15.16	0.062
2008	8.68	0.028	2016	16.92	0.066

若将表 9-2 中资料描绘成散点图,根据理论或经验可以做出判断,用双曲线来表示这两变量之间的关系是恰当的。双曲线的方程式是:$\frac{1}{y}=a+b\frac{1}{x}$

求解回归方程之前,先做变量变换,令$\frac{1}{y}=y'$,$\frac{1}{x}=x'$

则得 $y'=a=bx'$的直线形式,然后按照直线回归分析方法求解直线回归方程。所需计算数据如表 9-3 中列示。

表 9-3　例 7-2 计算表

年　份	x	y	x'	y'	$(x')^2$	$x'y'$
2001	5.23	0.016	0.19	62.50	0.036	11.88
2002	5.63	0.015	0.18	66.67	0.032	12.00
2003	5.94	0.016	0.17	62.50	0.028	10.63
2004	6.35	0.019	0.16	52.63	0.026	8.42
2005	6.88	0.025	0.15	40.00	0.023	6.00
2006	7.53	0.029	0.13	34.48	0.017	4.48
2007	7.96	0.028	0.13	35.71	0.017	4.64
2008	8.68	0.028	0.12	35.71	0.014	4.29
2009	9.35	0.031	0.11	32.25	0.012	3.55
2010	9.82	0.034	0.10	29.11	0.010	2.94
2011	10.63	0.034	0.09	29.41	0.008	2.65
2012	11.71	0.035	0.09	28.57	0.008	2.57
2013	13.06	0.044	0.08	22.73	0.006	1.82
2014	14.13	0.056	0.07	17.86	0.005	1.25
2015	15.16	0.062	0.07	16.12	0.005	1.13
2016	19.92	0.066	0.06	15.15	0.004	0.91
合计			1.90	581.71	0.251	79.16

计算过程如下:

$$b=\frac{n\sum x'y'-\sum x'\sum y'}{\sum x'^2-(\sum x')^2}=\frac{16\times(79.16)-1.9\times(581.71)}{16\times(0.251)-(1.9)^2}=397.32$$

$$a=\bar{y}'-b\bar{x}'=\frac{581.71}{16}-397.32\times\left(\frac{1.9}{16}\right)=-11.32$$

$$\therefore y'=-11.32+397.32x'\quad 即\frac{1}{y}=-11.32+397.32\frac{1}{x}$$

二、常用的非线性回归模型简介

(一) 解释变量可以直接替换的非线性回归模型

对于解释变量是非线性的,但参数之间是线性的模型,可以利用变量直接代换的方法将模型线性化。

1. 多项式函数模型

对于多项式模型

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + u$$

令

$$z_1 = x, \quad z_2 = x^2, \quad z_k = x^k$$

原模型可化为线性形式

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_k z_k + u$$

即可利用多元线性回归分析的方法处理。

这类模型广泛地用于生产和成本函数。例如总成本函数可表示为

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_i^2 + \beta_3 x_i^3$$

式中，y 为总成本；x 为产出。

2. 双曲线(倒数)模型

对于双曲线模型

$$y = \beta_0 + \beta_1 \frac{1}{x} + u$$

令 $Z=\frac{1}{x}$，原模型可化为线性形式

$$y = \beta_0 + \beta_1 Z + u$$

即可利用一元线性回归分析的方法处理。

在社会经济变量中，如工资变化率与失业率之间(菲利普斯曲线)、平均固定成本与产量之间(平均固定成本曲线)等都存在双曲线这种类型的依存关系。

3. 双对数函数模型和半对数函数模型

(1) 双对数函数模型。

对于双对数函数模型

$$\ln y = \beta_0 + \beta_1 \ln x + u$$

令

$$y^* = \ln y, \quad x^* = \ln x$$

于是原模型可化为标准线性模型

$$y^* = \beta_0 + \beta_1 x^* + u$$

变换后的模型不仅参数是线性的，而且变换后的解释变量也是线性的。

在实际工作中，双对数模型的应用非常广泛。其原因在于，由于回归线是一条直线(y 和 x 都是对数形式)，所以它的斜率(β_1)为一常数。对于这个模型，其斜率度量了 y 关于 x 的弹性(系数)，因为

$$\beta_1 = \frac{\mathrm{d}y^*}{\mathrm{d}x^*} = \frac{\mathrm{d}(\ln y)}{\mathrm{d}(\ln x)} = \frac{\Delta y/y}{\Delta x/x} = E$$

所以弹性为一常数。它表示 x 变动 1%，y 变动了 β_1%。由于这个特殊的性质，双对数模型又称为不变弹性模型。

弹性(如需求函数中的价格弹性、收入弹性，生产函数中的资金弹性、劳动力弹性等)是经济分析中的重要指标，如果所研究的经济现象能用双对数模型来描述，则参数估计后就可以直接利用回归系数进行弹性分析了。

当社会经济变量观测值的对数散点图，近似一条直线时，就可以双对数模型来描述或

拟合。

(2) 半对数函数模型。

我们把函数形式为 $\ln y=\beta_0+\beta_1 x+u$(对数-线性模拟)或 $y=\beta+\beta\ln x+u$(线性-对数模型)称为半对数模型。

对于对数-线性模型

$$\ln y = \beta_0 + \beta_1 x + u$$

$$\beta_1 = \frac{\mathrm{d}y^*}{\mathrm{d}x} = \frac{\mathrm{d}(\ln y)}{\mathrm{d}x} = \frac{\Delta y/y}{\Delta x}$$

$$\frac{\Delta y}{y} = \beta_1 \Delta x$$

它表示 x 变动一个单位,y 将变动 $\beta_1\%$的百分比。即 y 的相对变动百分比等于 β_1 乘以 x 的绝对变化量。

对于线性-对数模型 $y=\beta+\beta\ln x+u$

$$\beta_1 = \frac{\mathrm{d}y}{\mathrm{d}x^*} = \frac{\mathrm{d}y}{\mathrm{d}(\ln x)} = \frac{\Delta y}{\Delta x/x}$$

$$\Delta y = \beta_1 \frac{\Delta x}{x}$$

它表示 x 变动 1%,y 将变动 β_1 个单位的绝对量。即 y 的绝对变化量等于 β_1 乘以 x 的相对变化量。

半对数模型通常用于测度经济变量的增长率,如测定人口增长率、劳动力增长率、货币供应量增长率、GDP 增长率、商品需求量增长率、进出口贸易增长率等,所以半对数模型又称为增长模型。

4. 逻辑斯蒂(Logistic)曲线 $y=\frac{1}{a+be^{-x}}$

令 $y'=\frac{1}{y}$,$x'=\mathrm{e}^{-x}$

则有 $y'=a+bx'$

(二) 解释变量需间接替换的非线性回归模型

1. 指数曲线 $y=ab^x$

两边取对数得

$$\log y = \log a + x\log b$$

令 $\log y=y'$,$\log a=a'$,$\log b=b'$

则有 $y'=a+bx'$

2. 幂函数曲线 $y=dx^b$

两边取对数得

$$\log y = \log d + b\log x$$

令 $y'=\log y$,$x'=\log x$,$a=\log d$

则有 $y'=a+bx'$

3. 龚珀兹(Gompertz)曲线 $y=de^{bx}$

两边取对数得

$$\ln y = \ln d + bx$$

令 $y' = \ln y, a = \ln d$

则有

$$y' = a + bx'$$

三、几点说明

以上所述线性变换的方法具有简单易行的优点。但是，在实际应用时应注意以下几个问题。

（1）为了能够根据样本观测值，对通过变换得到的线性回归方程式进行估计，该方程中的所有变量都不允许包含未知的参数。这是因为 Y 包含了未知的参数是不可观测的。

（2）在以上的讨论中，我们省略了非线性回归方程中包含的随机误差项。但事实上与线性回归分析的场合一样，非线性回归分析也要考虑随机误差项的问题。只有当变换后新模型中包含的误差项能够满足各种标准假定时，新模型中回归系数最小二乘估计量的各种理想性质才能成立。

（3）严格地说，上述的各种线性变换方法只是适用于变量为非线性的函数。对于参数为非线性或参数与变量均为非线性的函数来说，即使有可能进行线性变换和回归估计，也无法得到原方程中非线性参数的无偏估计量。

（4）并不是所有的非线性函数都可以通过变换得到与原方程完全等价的线性方程。在遇到这种情况时，还需要利用其他一些方法如泰勒级数展开法等去进行估计。由于这些方法比较复杂，超出了本书的程度，这里不做进一步的介绍。

四、回归模型的优先问题

当有几个模型可供我们选择时，我们可用下列指标比做出判断。

（1）均方根误差（root mean squared error，RMSE），数值越小越好。

$$\text{RMSE} = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}$$

（2）平均绝对误差（mean absolute error，MAE），数值越小越好。

$$\text{MAE} = \frac{1}{n}\sum |y - \hat{y}|$$

（3）平均绝对百分比误差（mean absolute percent error，MAPE）。

$$\text{MAPE} = \frac{1}{n}\sum \left|\frac{y - \hat{y}}{y}\right| \times 100\%$$

一般认为，如果 $\text{MAPE} < 10$，则认为预测精度较高（数值越小越好）。

（4）希尔不等系数（theil inquality coefficient，Theil IC）。

$$\text{Theil IC} = \frac{\sqrt{\frac{1}{n}\sum (y - \hat{y})^2}}{\sqrt{\frac{1}{n}\sum \hat{y}^2} + \sqrt{\frac{1}{n}\sum y^2}}$$

希尔不等系数总是介于 0 和 1 之间，数值越小表明预测精度越高。

(5) 偏差率(bias proportion,BP)、方差率(variance proportion,VP)、协方差率(covariance proportion,CP)。

$$BP=\frac{(\bar{\hat{y}}-\bar{y})^2}{\sum(\bar{y}-y)^2/n},\quad VP=\frac{(\sigma_{\hat{y}}-\sigma_y)^2}{\sum(\hat{y}-y)^2/n},\quad CP=\frac{2(1-r)\sigma_{\hat{y}}\sigma_y{}^2}{\sum(\hat{y}-y)^2/n}$$

三个指标之和等于 1,三个指标的分母为均方误差。当预测比较理想时,均方误差大多集中在协方差率上,偏差率和方差率很小。

在使用 Eviews 软件的条件下,以上五个指标,在方程窗口单击 Forecast,即可得到。

(6) 修正的可决系数 R^2(Adjusted R Square 越大越好)。

(7) 对数似然值(log likelihood,L)越大越好。

$$L=-\frac{n}{2}\log 2\pi-\frac{n}{2}\log 2\hat{\sigma}-\frac{n}{2}$$

(8) 赤池信息准则(Akaike information criterion,AIC),越小越好。

$$AIC=-\frac{2L}{n}+\frac{2m}{n}\quad (m\text{ 为参数个数})$$

(9) 施瓦兹准则(Schwarz criterrion,SC),越小越好。

$$SC=-\frac{2L}{n}+\frac{m\ln n}{n}$$

在使用 Eviews 软件的条件下,以上四个指标一并出现在参数估计结果中。

第五节 Excel 相关回归分析应用

以例 9-1 资料为例,说明 Excel 在相关回归分析中的应用。

一、利用“粘贴函数”计算相关系数

首先建立数据文件如图 9-3 所示(B、C 两列)。

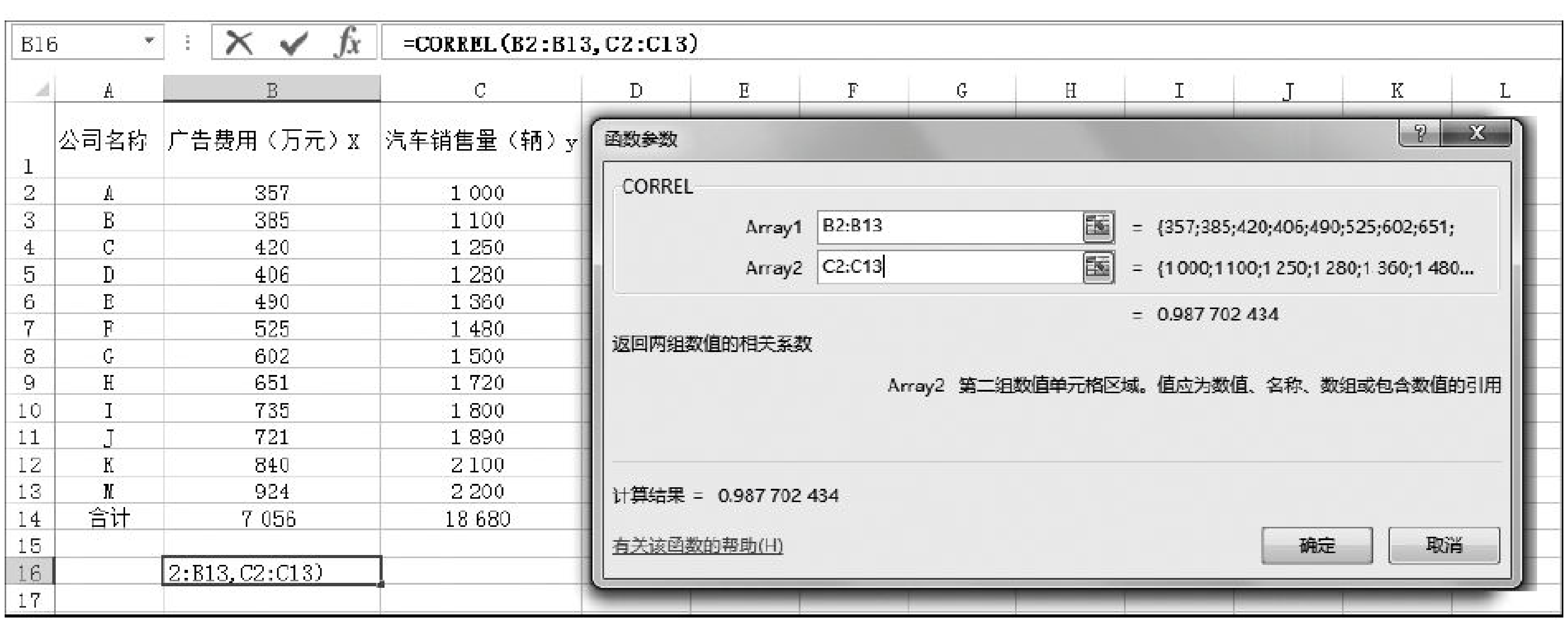


图 9-3 利用“粘贴函数”计算相关系数对话框

单击“公式”→“插入函数”→“统计函数”→“CORREL”,进行设置(图 9-3),再点击“确定”按钮,即可得到相关系数 0.987 702 434。

二、使用分析工具库计算相关系数

依次单击“数据”→“数据分析”→“相关系数”，进行设置（图 9-4），单击“确定”，即可得到相关系数 0.987 702 434。



图 9-4　使用分析工具库计算相关系数对话框

三、使用图表向导做相关图

依次点击“插入”→“推荐的图”→“散点图”，即可得到相关散点图（见图 9-5）。

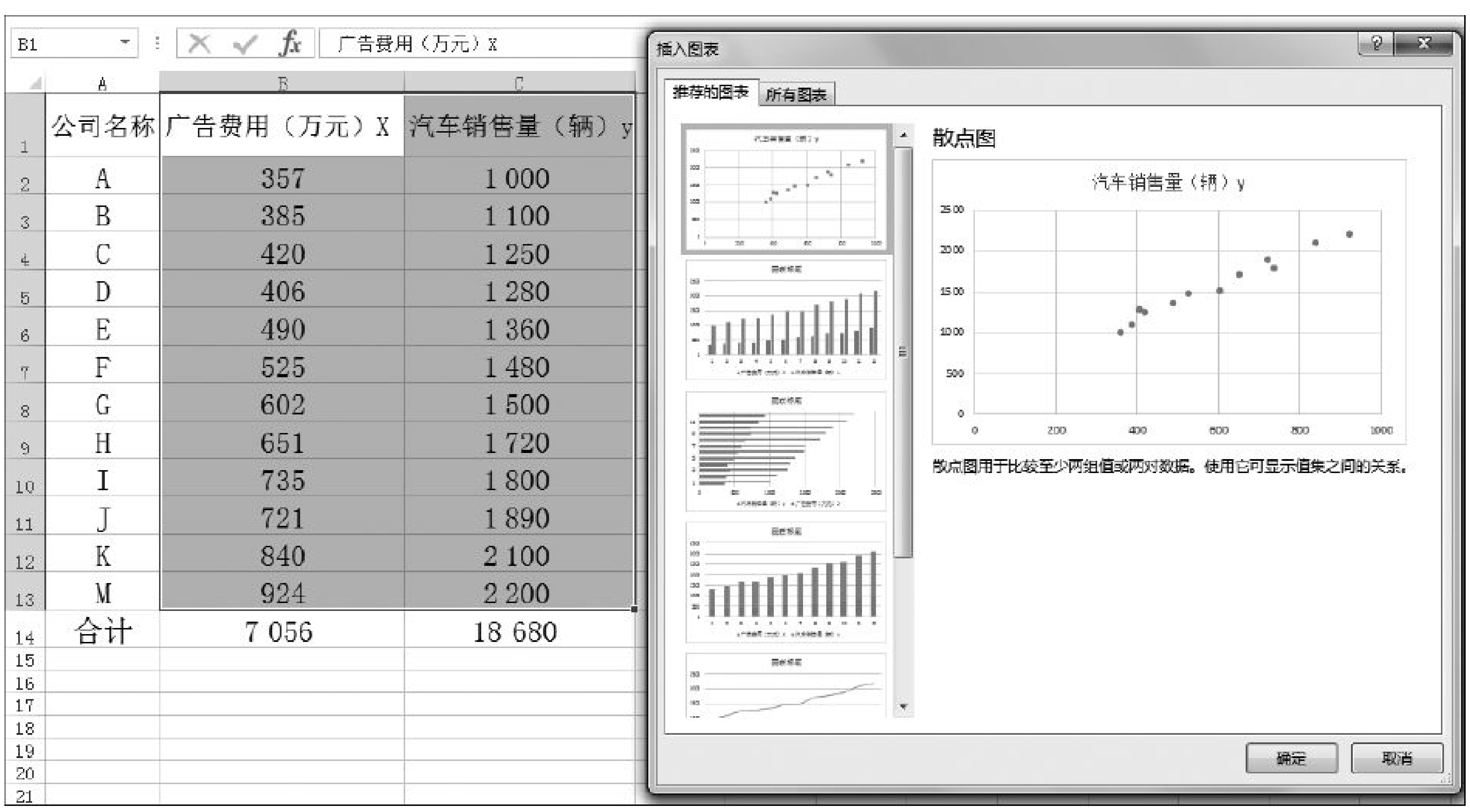


图 9-5　使用图表向导做相关图

四、利用分析工具库进行回归分析

依次单击“数据”→“数据分析”→“回归”，进行设置（图 9-6），单击“确定”按钮，即可得到运行结果，如图 9-7 所示。



图 9-6　回归分析设置对话框

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	回归统计						
4	Multiple	0.987 702					
5	R Square	0.975 556					
6	Adjusted	0.973 112					
7	标准误差	63.3 704					
8	观测值	12					
9							
10	方差分析						
11		df	SS	MS	F	gnificance F	
12	回归分析	1	1 602 709	1 602 709	399.1	2.17E-09	
13	残差	10	40 158.07	4 015.807			
14	总计	11	1 642 867				
15							
16		Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	363.6891	62.455 29	5.823 191	0.000 168	224.53	502.848 1
18	X Variabl	2.028 873	0.101 558	19.97 749	2.17E-09	1.802 588	2.255 159

图 9-7　回归分析输出结果

思　考　题

1. 简述相关关系、简单相关和复相关、线性相关和非线性相关、正相关和负相关。
2. 简述相关分析的主要内容。
3. 简述回归分析的主要内容。
4. 简述相关分析和回归分析的联系和区别。
5. 简述相关系数、复相关系数、简单相关系数、偏相关系数、样本决定系数的概念。
6. 简述一元线性回归模型的基本假定、多元线性回归模型的基本假定。
7. 什么是最小二乘法？
8. 简述随机误差项 u 的内容、σ_u^2 的无偏估计量。

9. 简述统计显著性检验思想。

10. 简述个别值预测和均值预测的概念。

练　习　题

1. 调查10位同学每天学习时间与高考成绩，得到表9-4所示资料。

表9-4　某校10位同学每天学习时间与高考成绩资料

学　生	学习时间/小时	学习成绩/分	学　生	学习时间/小时	学习成绩/分
学生1	4.0	60	学生6	7.0	90
学生2	5.0	72	学生7	7.5	91
学生3	5.5	80	学生8	8.0	95
学生4	6.0	86	学生9	8.5	96
学生5	6.5	88	学生10	9.0	96.5

要求：

(1) 计算学习时间与学习成绩之间的相关系数。

(2) 确定学习成绩对学习时间的直线回归方程。

(3) 根据建立的直线回归方程，设甲学生每天学习时间为13小时，请对他的成绩进行预测(置信度为95%)。

2. 某厂某产品上半年各月产量及单位成本资料如表9-5所示。

表9-5　某厂某产品上半年各月产量及单位成本资料

月份	产量/万件	成本/元/件
1	3.0	15.0
2	3.5	14.0
3	4.0	13.0
4	4.5	12.0
5	5.0	11.0
6	6.0	9.0

(1) 试确定直线回归方程，并指出产量每增加10 000件时，单位成本平均下降多少元？

(2) 建立线性回归模型，对产量为6.5万件时的单位成本进行区间估计(置信度为95%)。

3. 在其他条件不变的情况下，某种商品的需求量(y)与该商品的价格(J)有关。现对给定时期内的价格与需求量进行观察，得到如表9-6所示数据。

表9-6　某种商品的需求量与价格关系

价格 J	100	110	120	125	130	135	138	140	150	155
需求量 y	70	60	58	55	53	50	45	40	37	30

要求:

(1) 计算价格与需求量之间的相关系数。

(2) 拟合需求量对价格的回归直线,并解释回归系数的含义。

4. 某企业广告费和销售收入历史统计资料如表 9-7 所示。

表 9-7 某企业广告费和销售收入历史统计资料

广告费 x	1	2	3	4	5	6	7	8
销售收入 y	10	14	18	20	25	28	30	40

要求:

(1) 绘制散点图。

(2) x 与 y 之间是否大致呈线性关系?

(3) 用最小二乘法求出回归方程。

(4) 求回归标准误差。

(5) 回归系数是否显著?

(6) 计算 x 与 y 的决定系数。

第十章

统计分析报告

【统计知识 ABC】

[A]倍数与翻番

倍数和翻番都是描述数量增加的一种抽象化的表现形式，但二者的含义和计算方法截然不同。

倍数是由两个有联系的指标对比，将对比的基数抽象化为 1 而计算出来的相对数，常常用于比数（分子）远大于基数（分母）的场合。如某城市 2000 年人均住房使用面积达到 14.8 平方米，为 1978 年 3.8 平方米的 3.9 倍（14.8÷3.8＝3.9 倍）。再如，甲地人均国内生产总值 8 000 元，为乙地 3 000 元的 2.7 倍（8 000÷3 000＝2.7 倍）。

翻番是指数量加倍。如 1 变为 2(1×2)，2 变为 4(2×2)，3 变为 6(3×2)……A 变为 $A\times2$，翻两番为 $(A\times2)\times2=A\times2^2$，是指原基数在翻一番的基础上再翻一番。例如，党的十六大报告中提出的全面建设小康社会的目标，即在优化结构和提高效益的基础上，国内生产总值到 2020 年力争比 2000 年翻两番，综合国力和国际竞争力明显增强。也就是说 2020 年我国的国内生产总值将是 2000 年的 4 倍。

翻番和倍数有一定联系，$2n$ 即翻番后的水平相当于原基数的倍数，如原基数是 5，翻三番为 5×2^3，翻三番后达到的水平为 40，相当于基数 5 的 2^3 倍，即 8 倍。可见“番”是按几何级数计算的，“倍”是按算术级数计算的。

（资料来源：http://www.stats-tj.gov.cn/Article/ljtj/tjzswd/tjylhff/200612/5470.html.）

[B]百分数与百分点

百分数也称百分比，是相对指标最常用的一种表现形式。它是将对比的基数抽象化为 100 而计算出来的相对数，用“%”表示。它既可以表示数量的增加，也可以表示数量的减少。运用百分数时，要注意概念的精确。如“比过去增长 20%”，即过去为 100，现在是 120；比过去降低 20%，即过去是 100，现在是 80；“降低到原来的 20%”，即原来是 100，现在是 20。运用百分数时，还要注意有些数最多只能达到 100%，如产品合格率，种子发芽率等；有些百分数只能小于 100%，如粮食出粉率等；有些百分数可以超过 100%，如毛入学率等。要注意区别占、超、为、增的含义和用法，“占计划百分之几”指完成计划的百分之几；“超计划的百分之几”，就应该扣除原来的基数；“为去年的百分之几”就是等于或相当于去年的百分之几；“比去年增长百分之几”应扣掉原有的基数。

百分点是指不同时期以百分数形式表示的相对指标,如速度、指数、构成等的变动幅度,它是分析百分比增减变动的一种表现形式。例如,工业增加值今年的增长速度为19%,去年的增长速度为16%,今年比去年的增长幅度提高了3个百分点(19－16)。今年物价上升了8%,去年物价上升了10%,今年比去年物价上升幅度下降了2个百分点(8－10)。

百分数是两期数据相除(或再减1),表示事物的发展速度(或增长速度),百分点是两个百分数(去掉百分号)相减,表示事物的增长量,是百分数的增长量(比如比重等)。

百分数和百分点有严格的区别,使用时要注意。下面举三个例子予以说明。

例1:某地区生产总值,1992年为100亿,其中第一产业贡献23.8亿,占23.8%;1993年为125亿,其中第一产业贡献26.5亿,占21.2%。我们可以说:该地区生产总值中,第一产业占的比重,1993年比1992年下降2.6个(21.2－23.8＝－2.6)百分点,但不能说下降2.6%。还可以说:第一产业增加值1993年比1992年增长11.34%[即(26.5/23.8－1)×100%≈11.34%],但不能说增长11.34个百分点。具体怎样使用,看要说明什么问题而定。

例2:某企业计划2016年比2015年利润增加9.8%,实际增加10%。我们可以说:该企业2016年实际利润比计划高了0.2个(10－9.8＝0.2)百分点;或者说:某企业2015年实际利润比计划多完成2.04%[即(10%/9.8%－1)×100%≈2.04%]。

例3:2015年,国家财政性教育经费支出占GDP的比例为4.26%,2014年为4.10%。这样2015年国家财政性教育经费支出占GDP的比例比2014年增加了0.16个百分点,不能说增加了0.16%。实际上,2015年国家财政性教育经费支出占GDP的比例比2014年增加了3.90%[即(4.26%/4.10%－1)×100%≈3.90%。

另外,对时间数列进行分析时,百分数涉及绝对数时间数列的两期数据,百分点涉及绝对数时间数列的三期数据。即只要有两期总量指标,就可以用百分数比较分析,而百分点就必须有三期总量指标方可进行。

例:某地区农民人均纯收入,2015年为3 400元,2016年为4 080元,2017年为4 500元。依据这些数据,可以进行如下分析:该地区农民人均纯收入2016年比2015年增长20%[即(4 080/3 400－1)×100%＝20%],2017年比2016年增长10.29%[即(4 500/4 080－1)×100%≈10.29%];该地区农民人均纯收入增长速度2017年比2016年降低9.71个百分点。从以上计算可以看出,百分数依据两期数据可以得出,而百分点涉及了三期数据。

(资料来源:http://blog.sina.com.cn/s/blog_4e88c7af01000ahu.html.)

[C]现行价格与可比价格

现行价格就是当年价格,是指报告期当年的实际价格,如工业品的出厂价格、农产品的收购价格、商品的零售价格等。用当年价格计算的一些以货币表现的物量指标,如国内生产总值、工业总产值、农业总产值、农副产品收购总额和社会商品零售总额等,反映当年的实际情况,使国民经济指标互相衔接,便于考察社会经济效益,便于对生产、流通、分配、消费之间进行综合平衡。因此,当我们需要反映当年的实际收入情况时就应采用当年价

格。例如：2016 年我国国内生产总值为 744 127 亿元，它反映 2016 年在我国领土范围内所生产的以货币表现的产品和劳务总量。

按当年价格计算的以货币表现的指标，在不同年份之间进行对比时，因为包含各年间价格变动的因素，不能确切地反映实物量的增减变动，必须消除价格变动的因素后，才能真实地反映经济发展动态。因此，在计算增长速度时，一般都使用可比价格计算。例如，我们要计算 2015 年工业总产值增长速度，因为用当年价格表示的 2015 年、2014 年工业总产值存在着价格变动因素，因此不能直接用来计算增长速度，而应采用消除了价格因素后的可比价格进行计算。2015 年工业总产值按当年价格计算为 235 183.5 亿元，按 2010 年不变价格为 242 081.8 亿元，2014 年则分别为 233 856.4 亿元和 228 420.8 亿元，如按当年价格计算，2015 年比 2014 年增长速度为(235 183.5÷233 856.4－1)×100％＝0.57％，但由于没有剔除价格变动因素的影响，故不能确切地反映工业生产实物量的增长状况，而按可比价格计算的增长速度则为(242 081.8÷228 420.8－1)×100％＝5.98％，这一速度就较为确切地反映出工业生产实物量的增长。

不变价格，也叫固定价格，它是用某一时期同类产品的平均价格作为固定价格来计算各个时期的产品价值，目的是消除各时期价格变动的影响，保证前后时期之间、地区之间、计划与实际之间指标的可比性。

我国自开始核算国内生产总值以来，共有 1952 年、1957 年、1970 年、1980 年、1990 年、2000 年、2005 年、2010 年 8 个不变价基期，目前的基期是 2010 年。也就是说，2011—2015 年的不变价国内生产总值是按照 2010 年价格计算的。

（资料来源：http://www.stats.gov.cn/tjzs/tjcd/t20020528_20656.html.）

第一节　统计分析报告概述

一、统计分析报告的含义

所谓统计分析是运用统计方法对客观现象的状态、规律、前景进行认知的一种活动。

统计分析报告是根据统计学的原理和方法，运用大量统计数据来反映、研究和分析社会经济活动的现状、成因、本质和规律，并做出结论，提出解决问题办法的一种统计应用文体。

统计分析报告是统计分析过程中所形成的论点、论据、结论的集中表现。它不同于一般的总结报告、议论文、叙述文和说明文，更不同于小说、诗歌和散文，它是运用统计资料和统计方法、数字与文字相结合，对客观事物进行分析研究结果的表现。统计分析结果可以通过表格、图形和文章等多种形式表现出来。文章是统计分析报告的主要形式，它是所有表现形式中最完善的，这种形式可以综合而灵活地运用表格、图形等形式，可以表现出单纯表格形式、图形形式难以充分表现的情况，可以使分析结果鲜明、生动、具体，可以进行深刻的定性分析。因此，统计分析报告是对研究过程进行表述的文章，是统计分析结果的最终形式。

二、统计分析报告的特点

统计分析报告的特点有以下几点。

(一)统计分析报告以统计数据为主体

统计分析报告主要以统计数字语言,来直观地反映事物之间的各种复杂的联系,以确凿的数据来说明具体时间、地点、条件下社会经济领域的成就和经验、问题与教训、各种矛盾及其解决办法。它不同于用艺术形象刻画的文艺作品,也不同于旁征博引进行探讨研究的各种论文,而是以统计数字为主体,用简洁的文字来分析叙述事物量的方面及其关系,并进行定量分析。

(二)统计分析报告以统计分析为基础

统计分析报告运用一整套统计特有的科学分析方法(如对比分析法、动态分析法、因素分析法、统计推断等),结合统计指标体系,全面、深刻地研究和分析社会经济现象的发展变化。同时,利用统计部门的优势,从数量方面来表现事物的规模、水平、构成、速度、质量、效益等情况,并把定量分析与定性分析结合起来。

(三)统计分析报告具有很强的针对性和实用性

统计分析报告一般针对各级党政领导和社会各界普遍关心的难点、热点、焦点问题进行分析,有的放矢,针对性强。

统计分析报告不但包含了统计数据反映的信息,更为重要的是,它还能进行分析研究,能进行预测,能指出工作中的不足和问题,能提出有益于今后工作的措施和建议,从而直接满足党政领导和社会各界在了解形势、制定政策、编制计划、经营管理、检查监督、总结评比、科研教学等方面的实际需要。

(四)统计分析报告注重准确性和时效性

准确是统计分析报告乃至整个统计工作的生命。统计分析报告的准确性除了数字准确,不能有丝毫差错,情况真实,不能有虚假之外,还要求论述有理,不能违反逻辑;观点正确,不能出现谬误;建议可行,不能脱离实际。

同时,统计分析报告具有很强的时效性。失去了时效性,也就失去了实用性,统计分析报告写得再好也成了无效劳动。因此,统计分析报告要提供在领导决策之前和社会各界需要之时。

(五)统计分析报告具有独特的表达方式和结构

统计分析报告属于应用文体中的说明文,基本表达方式是以事实来叙述,让数字说话,在阐述中议论,在议论中分析。在表现事物时,不是用夸张、虚构、想象等手法,而是用较少的文字、精确的数据,言简意赅、精练准确地表达丰富的内涵。

统计分析报告在结构上的突出特点是脉络清晰、层次分明。一般是先摆数据、事实，进行各种科学的分析，进而揭明问题，亮出观点，最后有针对性地提出建议、办法和措施。统计分析报告的行文，通常是先后有序，主次分明，详略得当，联系紧密，做到统计资料与基本观点统一，结构形式与文章内容统计，数据、情况、问题和建议融为一体。

三、统计分析报告的作用

统计分析报告的作用有以下几点。

（一）统计分析报告是衡量统计工作水平的综合标准

统计分析报告是统计工作的最终成果。在一定意义上也就是统计设计、统计调查、统计整理、统计分析与统计分析写作全部工作水平的综合。前面几个环节是统计的基础工作，统计分析才是出成果的阶段。一般来说，高质量的统计分析报告，来自高质量的统计设计、统计调查、统计整理、统计分析和统计分析写作。但是，如果仅有较好的写作水平，统计设计、统计调查、统计整理和统计分析都是低质量的，也不可能产生高质量的统计分析报告。因此，统计分析报告写不好，当然是统计工作水平不高的表现。更重要的是，还要具备方方面面的科学文化知识，包括统计专业知识。需要掌握党和国家的方针政策，需要具备较强的观察能力、思维能力、创新能力、组织能力等。所以，统计分析报告的质量如何也就反映了统计工作水平如何，这是一个非常重要的综合标准。另外，统计分析的结果虽可以用多种形式表达（如表格式、图形式、文章式等），但只有文章形式的统计分析报告为最好，也最为常用。因此，统计分析报告也是表现统计成果的好形式。

（二）统计分析报告是传播统计信息的有效工具

现代社会是信息时代，信息已成为重要资源。统计信息又是社会信息的主体，而且是最全面、最稳定、较准确的信息。统计信息要通过载体传播，而统计分析报告是主要载体之一，适合在报纸杂志上发表，传播条件比较简便，具有较大的信息覆盖面，是传播统计信息的有效工具。

（三）统计分析报告是党政领导和社会各界决策的重要依据

现代社会经济管理必须科学决策，而科学的决策又必须依据准确、真实的统计数据。统计分析报告把原始资料信息加工成决策信息，它比一般的统计资料更能深入地反映客观实际，更便于党政领导和社会各界接受利用。因而，统计分析报告是党政领导和社会各界决策的重要依据。

（四）统计分析报告是统计服务与统计监督的主要手段

统计分析报告把数据、情况、问题、建议等融为一体，既有定量分析又有定性分析，比一般的统计数据更集中、更系统、更鲜明、更生动地反映了客观实际，又便于人们阅读、理解和利用，是表现统计成果的好形式与传播统计信息的有效工具，自然也就成了统计服务

与统计监督的主要手段。

(五)统计分析报告是增进社会了解、提高统计社会地位的主要窗口

由于历史、体制等原因,一般大众缺乏统计知识,对统计不够了解,对统计工作不够重视,认为“统计是三分统计,七分估计”,“统计工作只是加加减减,填个表而已”,把统计置于可有可无的地位。要改变这种状况,一方面要加强统计宣传工作,扩大统计的影响,提高人们的认识;另一方面则要提高统计工作水平,写好统计分析报告,做好统计服务和统计监督工作,提高统计工作的社会地位。

(六)统计分析报告有利于促进统计工作自身的发展

统计分析报告的质量,反映了统计工作的水平。在统计分析报告的写作过程中,能有效地检验统计工作各个环节的工作质量,发现问题及时改进,使统计工作得到改善、加强和提高。另外,经常撰写统计分析报告,能综合锻炼提高写作人员的素质,全面增长统计人员的才干。

总之,统计分析报告十分重要,那种认为“统计报表是硬任务,统计分析是软任务”的说法,是完全错误的,是万万要不得的。

四、统计分析报告的种类

统计分析报告可以从不同角度来划分种类。

(一)按统计分析报告的内容和作用不同,可以分为统计公报、进度统计分析报告、综合统计分析报告等

1. 统计公报

统计公报是政府统计机构通过报刊向社会公众公布一个年度国民经济和社会发展情况的统计分析报告。一般是由国家、省一级以及计划单列的省辖市一级的统计局发布的,如《国家统计局关于 2012 年国民经济和社会发展统计公报》。

2. 进度统计分析报告

进度统计分析报告主要以定期报表为依据,反映社会经济的发展情况,分析其影响和形成的原因,如月度分析、季度分析和年度分析。从时间上看,它可分为定期和不定期的、期中的和期末的统计分析报告;从内容上看,它又可分为专题和综合统计分析报告两种。

进度统计分析报告必须讲究时效,力求内容短小精悍,结构简单规范,看后一目了然。

3. 综合统计分析报告

综合统计分析报告是从客观的角度,利用大量丰富的统计资料,对国民经济和社会发展的规模、水平、结构和比例关系、经济效益以及发展变化状况,进行综合分析研究所形成的一种统计分析报告。

4. 专题统计分析报告

专题统计分析报告是对社会经济现象的某一方面或某一问题进行专门的、深入研究的一种分析报告。它的目标集中,内容单一,不像综合分析报告那样,要反映事物的全貌。

正因为如此，专题统计分析报告更要求突破时间和空间的限制，根据领导和社会公众的需要灵活选题，做到重点突出、认识深刻。

5. 典型调查报告

典型调查报告是根据调查的目的、要求，有意识地选择少数有代表性的单位进行深入实际调查后所写成的报告。深入实际，进行调查研究，是各级领导、各部门了解情况、指导工作经常采用的一种工作方法。习惯上称为"解剖麻雀"，统计上叫作典型调查。

（二）按写作类型不同，可分为说明型、快报型、计划型统计分析报告等

1. 说明型统计分析报告

这是对统计报表进行说明的统计分析报告，亦称为"文字说明"，也就是我们通常所说的报表说明。这种说明，主要是对报表的数据做文字的补充叙述，配合报表进一步反映社会经济情况。这种补充叙述主要是针对报表中某些变化较大的统计数字，这也可以帮助本单位领导审查报表，以保证数字的质量。这是说明型统计分析报告的基本作用。

严格地说，说明型统计分析报告，因为它只是附属统计报表，所以不能独立成篇，也无完整的文章形式。但由于它也具备统计分析报告的基本特点，我们可以把它看成统计分析报告的雏形。

2. 快报型统计分析报告

这是一种期限短、反映快的统计分析报告。一般按日、周、旬、半月写作的定期统计分析报告。快报型统计分析报告的突出特点是一个"快"字。按日写作的统计分析报告，常在第二天上午上班不久就要递交主管领导，其他以此类推。由于这种快的特点，快报型统计分析报告常用于反映生产进度、工程进度等，便于管理层了解情况，对生产和工作进行及时指导，所以快报型统计分析报告在企业用得比较普遍。

3. 计划型统计分析报告

这是检查计划执行情况的统计分析报告，按月、季、半年和年度检查计划执行情况的定期统计分析报告都属于这种类型。

4. 总结型统计分析报告

这是对一定时期社会经济发展情况进行总结分析的统计分析报告。通过分析总结，可以全面认识一个地区、部门或单位的社会经济形势，或某个方面的情况，以便发扬成绩，总结经验教训，制订新的措施，为今后工作创造更好的条件。

总结型统计分析报告大多是半年、一年或三五年的统计分析报告。从内容上看，有综合总结、部门总结及专题总结。综合总结，是对地区的整个社会经济或企业整个生产经营的总结；部门总结，是对部门经济（农业、工业、商业）或某个车间的总结；专题总结，是对某些方面进行的专题总结。

5. 公报型统计分析报告

这是政府统计机关向社会公告重大社会经济情况的统计分析报告。统计公报是政府的一种文件，一般应由级别较高的统计机关发布。级别较低的统计机关不宜发表公报，但是可以采用统计公报的写作形式公布本地的社会经济发展情况，也应列入公报型。

6. 调查型统计分析报告

这是通过非全面的专门调查来反映部分单位社会经济情况的统计分析报告。其基本特点是只反映部分单位的社会经济情况,一般不直接反映和推断总体情况。而且它的资料和情况来源于非全面调查(抽样调查、重点调查和典型调查等),并不主要来自全面统计调查。

7. 分析型统计分析报告

这是通过分析着重反映社会经济现象具体状态的统计分析报告。它同调查型的主要区别是它既反映部分单位的情况,也反映总体的情况,并以总体情况为主。而且它的资料和情况来源是多方面的,可以是部分单位的调查资料,也可以是全面统计报表资料、历史资料的横向对比资料等,其中又以全面统计中的报表资料较多。目前,统计人员写作的统计分析报告,大多属于这种类型。

8. 研究型统计分析报告

这是着重研究解决问题办法和进行理论探讨的统计分析报告。它同分析型的统计分析报告的主要区别是:分析型对社会现象的认识仍停留在具体状态,而研究型则是对具体的状态上升理论的高度,提出理论性的见解或新的观点。所以,研究型比分析型的意义又进一步,是一种高层次的统计分析报告。

9. 预测型统计分析报告

这是评估社会经济发展前景的统计分析报告。它与研究型统计分析报告的主要区别是:研究型着重对趋势性、规律性进行定型研究,而预测型是在认识趋势及规律的基础上,着重对前景进行具体的定性和定量的研究。通过预测,人们可以超前认识社会经济发展前景,对制定方针、发展策略、编制计划、搞好管理具有很大的帮助。因此,预测型分析报告的作用很大,也属于高层次的统计分析报告。

10. 资料型统计分析报告

这是着重提供统计资料的统计分析报告。主要有两种形式:一是数字式(也就是数字文字化)。数字式虽以数字资料为主,但它有文章的形式,也有观点和简要的分析。历史资料、年报资料、横向对比资料等,都可以写成这种统计分析报告。二是概况式。例如,地方概况、部门概况、行业概况、企业概况等。这种概况式资料,是通过数字和文字提供简要而全面的基本情况,以使读者对某个地方、某个部门、某个行业或某个企业单位的概貌有所了解。

11. 信息型统计分析报告

这是以信息方式反映社会经济情况的统计分析报告。它有内容简要、篇幅短小、传递快速、读者面广等特点。信息型统计分析报告不只在报刊上发表,也可以写成党政领导的内部参阅材料,自治区统计局编发的“经济要情”和自治区“重大信息”就属于信息型的统计分析报告。

12. 综合型统计分析报告

这是综合多项内容的统计分析报告。有情况、有分析、有预测、有建议等多项内容。

13. 系列型统计分析报告

这是运用系列形式写作的一组统计分析报告。常用于反映和研究范围较广、层次较多、情况较复杂,又很重要的社会经济问题。

此外,按统计领域不同,又可分为工业、农业、商业、科技、教育、文化、卫生、体育、人

口、财政、金融、政法、人民生活、国民经济综合、核算等统计分析报告。按写作对象的层次划分，可分为微观、中观和宏观统计分析报告。按照时间长度分，可分为定期与不定期的统计分析报告。定期统计分析报告，一般是利用当年的定期统计报表制度的统计资料来定期研究和反映社会经济情况。根据期限不同，定期统计分析报告又可分为日、周、旬、半月、月度、季度、上半年、年度等统计分析报告。不定期的统计分析报告，主要是用于研究和反映不需要经常性定期调查的社会经济情况。

第二节　统计分析报告的撰写

一、统计分析报告的质量要求

统计分析报告的质量好坏，一般从两个方面来衡量：一是统计分析报告的深度和广度。即报告的内容是否丰富，对资料的分析和写作技巧如何。二是统计分析报告的时效性及产生的社会影响。即分析报告在实际工作中发挥的作用如何，也就是它的社会效益。后者是衡量分析报告质量的主要标准。国家及地方统计局每年组织评选优秀统计分析报告时提出四条评比标准，即基本质量要求如下。

（1）选题准确，能够紧密结合经济形势，配合党的中心任务，反映方针、政策的执行情况和效果，对党政领导的决策能起积极的作用。

（2）资料可靠，观点鲜明，分析深刻，提出一定的见解。

（3）时效性强，反映情况及时。

（4）主题突出，结构严谨，条理清晰，文字简洁。

这四条标准可概括为统计分析报告的“四性”，即准确性、针对性、时效性、逻辑性。当然，要写出一篇高质量的统计分析报告，还应在求“新”和求“深”上下功夫。

所谓“新”，是指创新。不仅内容有新意，形式也要新颖。要有所创新，就要树立新观念，研究新课题，挖掘新事物、新思想，选择新视角，反映新情况、新特点、新动态，写出新成就、新问题，分析新原因，总结新经验，提出新建议。所谓“深”，是指深入透彻，要掌握丰富的资料，进行深入的分析，以达到对研究对象有深刻、透彻的认识。

二、统计分析报告的选题

（一）选题的意义和原则

统计分析报告的写作，首先要解决写什么题目的问题，确定题目对统计分析报告写作很重要，①它关系到统计分析报告是否具有实用性，是否“产品对路”。如果没有实用性，写得再好也是不会有人需要的。②关系到写作过程是否能顺利进行。如果选择的题目难度超过了作者本身的能力和条件，写作也不成功，不但不能实现写作目的，而且会造成人力、物力、财力和时间的浪费。人们常说，“选好了题目就成功了一半”，这句话是很有道理的。

统计分析报告的题目有三种：一是任务题，它是领导交办或上级布置的题目；二是固定题，这是结合定期报表制度进行分析的题目，这种题目一般不变化；三是自选题，这是作者自己选择的题目。所以，我们所说的选题就是针对自选题而言的。

统计分析报告的选题应同时具备两个基本条件：一是有实用价值；二是有新颖性，必须有新的内容或新的见解。

选择题目应该遵循以下几条原则。

(1) 选具有现实意义的课题，或是与中心工作、全局性工作有密切联系的课题。

(2) 选国民经济发展中带有苗头性、动向性、突发性的问题。

(3) 选改革开放和社会主义现代化建设中出现的新情况、新问题、新经验。

(4) 选各方面有不同看法的重大问题。

(5) 选配合中心工作、重要会议提供材料的课题。

一般情况下，最好是结合自己的专业工作，选择自己熟悉的适合自己业务水平的、各项资料也比较齐全的课题来写。这样，成功的把握较大。切不可好高骛远，选题过大过难，以致至力不从心，半途而废，即使勉强写出来了，也不会有较好的质量。

(二) 选题的方法

统计分析报告的课题虽然很多，但不等于随便什么都可以写，而是抓住党政领导和社会各界需要知道了解他们尚未认识或充分认识的社会经济情况。这是主观与客观应该结合的点，常常表现为“注意点”“矛盾点”和“发生点”。统计分析报告的选题应该抓住这“三点”。

所谓“注意点”，就是党政领导和社会各界比较关注的热点问题。比较“物价上涨过快”“食品安全”“看病难、看病贵”等问题都是人们比较关注的社会热点问题。所谓“矛盾点”就是问题比较集中，影响比较大，争议比较多，但长期得不到很好解决的社会难点问题。比如“三农问题”“房价过高过快”“收入分配”问题等。所谓“发生点”，就是我们常的新情况、新问题、新联系和新趋势。比如社会主义新农村建设问题，新型城镇化问题，等等。抓住“发生点”来写作统计分析报告，意义是很大的。

那么在实际工作中如何才能实现“注意点”“矛盾点”和“发生点”呢？可以采取以下一些方法。

(1) 经常深入实际、深入群众、深入生产第一线。只有经常下到基层去，下到实际工作中去，才能掌握丰富、生动、真实、具体的第一手材料，就可以发现问题、研究问题。这样脑子里积累的问题多了，材料多了，写起文章来就会深刻得多，不至于枯燥贫乏、空洞无力。

(2) 经常了解国家及地方政策制定部门的意图和工作动向。主要是多关注有关部门的一些会议，如党的“十八大”，深入研读分析必要的文件。

(3) 经常研究统计资料。如统计报表、统计年鉴、统计历史资料等，因为其包含了丰富的社会经济信息，能比较全面反映社会经济活动的过程，并能暴露一些问题，因此细心地研究这些资料，并有意识地进行一些纵向、横向比较，并注意剖析其中的内部结构以及各种联系的变化，这些统计数据也往往能提示社会经济活动的“注意点”“矛盾点”和“发生点”。

(4) 加强理论学习，经常阅读报纸杂志。

(三) 选题的内容

在写作统计分析报告时，可以参考以下内容来选题。

(1) 围绕方针政策选题。可以从以下几个方面来选择题目：①研究社会经济发展中的

新苗头、新动向和新情况，为制定新的政策提供依据；②研究政策贯彻执行情况，反映新成就、新经验；③研究政策执行中的新问题，分析原因，提出建议，为检验和校正政策提供依据。

(2) 围绕中心工作选题。所谓中心工作，就是国家在一段时间内集中力量开展的某项工作。应该看到，在不同时期、不同地区、不同部门和单位，其中心工作是不同的，比如转变经济发展方式是当前比较大的中心工作。

(3) 围绕重点选题。所谓重点，就是在全局中处于举足轻重地位的某些部位或某项工作。

(4) 围绕经济效益选题。提高经济效益是经济发展的重点问题，应当作为写作统计分析报告的经常课题。

(5) 围绕民生选题。民生问题，简单地说，就是与百姓生活密切相关的问题，最主要表现在吃穿住行、养老就医子女教育等生活必需上面。民生问题也是人民群众最关心、最直接、最现实的利益问题。关注民生、重视民生、保障民生、改善民生，同党的性质、宗旨和目标一脉相承。教育是民生之基，就业是民生之本，收入分配是民生之源，社会保障是民生之安全网。这四大问题都是民生的基本问题。

(6) 围绕民意选题。对党和政府的方针政策，出台的一些重大问题的看法和意见，真实地表达人民群众的意向和要求。

三、统计分析报告的写作步骤

统计分析报告的写作步骤如下。

（一）拟定提纲

统计分析报告应在明确主题的基础上，提出说明主题的各种观点和依据，说明形成主题观点的主要因素有哪些，与主题相矛盾的有哪些，最后提出解决问题的办法(建议)。应形成以主题思路为主导的完整的、有结构的、分层的、较为具体的统计分析报告提纲框架。

提炼主题观点的常用方法有以下几种。

1. 演绎归纳法

从一般到特殊是演绎，从特殊到一般是归纳。在写统计分析时，凡是先下结论后摆材料的叫演绎法，如现状(结论)→原因→建议；先摆材料后根据材料而下结论的叫归纳法，即分析材料→得出结论。在一篇分析报告中，要灵活运用以上两种方法，全篇可以用演绎法，分段可以运用归纳法。

2. 纵横对比法

在统计上把时间连续叫作“纵”，把地区联系叫作“横”。要说明发展情况就必须和去年“同期”比，和各个历史时期比，这就是纵的对比法；要说明各单位某一指标的高低，就必须把各单位和各单位进行对比，这就是横的对比法。横的对比法能起到“泼点凉水”的清醒剂作用，使领导居安思危，采取积极措施，克服弱点，赶上先进地区。

采用纵横对比法时要注意可比性，一是要注意时间、地点、条件的可比性；二是要注意统计指标的计算口径、范围、计算方法要有可比性。

3. 层层剖析法

层层剖析法也叫层层剥笋法，就是对事物要寻根问底，把问题弄个水落石出。即皮→肉→骨→骨髓。道理简单，但要熟练运用是件极不容易的事。

4. 相互联系法

相互联系法是把相关的资料或指标联系起来进行综合分析。如工农业比例关系、国民生产总值中一、二、三产业之间的比例关系问题、积累与消费的关系问题等。把所掌握的资料与客观实际情况或者生活情况联系起来。要把所分析的问题与理论数据或者经验数据联系起来分析。如人口出生率中新生婴儿的男女性别比为 105∶100,在国内生产总值中投资约占 GDP 的 30%以内,财政收入占 GDP 的比重等,联系党在各个时期的方针政策进行分析。

5. 沙里淘金法

沙里淘金法主要解决如何选用最多的材料。材料用得好,用得恰当,就能起到观点和材料的统一,即水乳交融;材料选得不好,就很难说明观点,像油水分离,甚至造成观点和材料的矛盾。有时两个相似的材料都很好,我们也只能忍痛割爱,舍弃其中一个。

(二)准备资料

从内容上看,需要准备反映和表现主题特征的材料,形成主题特征的各种要素材料,以及显示主题一般特征的典型材料。

从类别上看,统计分析报告根据撰写统计分析报告所需收集的材料有如下几种。

1. 统计资料

统计资料是写作统计分析报告用得最多,也是最主要的材料,可分为以下几种材料。

(1) 定期报表资料:这主要指当年的定期报表数字资料,也包括定期的原始记录资料。

(2) 一次性调查资料:这里指统计普查抽样调查、重点调查典型调查的数字资料。

(3) 统计整理资料:这里主要指历史统计资料和统计年鉴资料,仍是数字资料。

(4) 统计分析资料:这是指已经印发的各种统计分析素材及统计分析报告。

(5) 统计图表资料:这是指各种形式的统计图、统计表。

(6) 统计书刊:这是指统计部门编印的有统计资料内容的书刊,如《统计年鉴》《中国统计》《统计研究》《统计日报》等。

2. 调查资料

调查资料是在特定的统计调查中所取得的情况或资料,是指在统计报表之外的,尚未写成统计分析资料的情况。主要由观察法、访谈法、问卷法、座谈法等获取。

3. 业务材料

业务材料是反映社会经济有关业务活动情况的文字材料。这些材料大多来自各业务部门以及有关的业务会议。比如财政会议、经济工作会议等。

4. 见闻材料

见闻材料是通过非统计调查的日常见闻所取得的活情况。这种活情况,一是指有文字记载的,如报刊发表的一些社会现象;二是指没有文字记载的是作者耳闻目睹的某些社会现象(如在街上发生的一些突发事件等),这些并非特意调查的见闻,有时也成为统计分析报告的材料。

5. 政策法规

政策法规是党和政府的有关方针、政策、法律、条例、规定、决定、决议等文件材料。

6. 有关言论

有关言论是革命导师、领袖、党政领导、古今中外的专家、学者的有关言论。这也是统

计分析报告论事说理的重要材料。

7. 书籍材料

书籍材料是有关的教科书、论著、专著、资料书与参考性的工具书等。在书籍材料中，主要是理论材料。

8. 报刊材料

报刊材料是报纸、期刊发表的各类材料，其中包括内部的、不定期的报刊材料。

9. 横向材料

横向材料是指同类地区以及市际、省际、国际之间的材料。有了这种材料，在写作统计分析报告时才便于进行横向比较。

此外，还要掌握一些必要的文字材料，如诗歌、成语、典故、谚语等。这些文字材料若在写作中运用得好，必能增加统计分析报告的生动性与可读性。

（三）确定统计分析报告的结构

所谓结构，就是文章的内部组织、内部构造，是对文章内容进行安排的形式。统计分析报告的结构，在过去有个俗成约定的格式，就是“一情况、二问题、三建议”这种三段式。还有一种就是：提出问题→分析问题→解决问题。这是最常见的也是经常用的两种格式，但统计分析报告的格式应该是多样化的。例如：有的统计分析报告是情况、问题、根源、预测、建议五个部分组成。有的是情况、问题、根源、建议四个部分组成。有的虽然也是三段式，但组成部分是情况、问题、根源或者是问题、根源、建议。还有的是两部分：情况、问题，或问题、根源，或问题、建议，或情况、建议。也有的则专门写情况，或专门写问题，或专门写建议。总之统计分析报告的结构应该不局限于三段式，应该是多种形式。

1. 结构的形式

统计分析报告的常用结构形式主要有以下三点。

（1）递进结构。按照事物的因果关系、逻辑关系展开。报告各部分之间一层一层，层层深入地衔接。例如，现状—原因—结果，现状—问题—对策，历史—现状—未来。

（2）并列结构。将所要表述的情况，分成并列的几个部分。如分析进口商品检疫状况，分国别美国、日本、德国等分别进行叙述。

（3）序时结构。按照事物发展的经过和时间先后进行表述。如改革开放前、改革开放后等。序时结构要注意精心取材，突出重点，切忌平铺直叙。

2. 标题

标题也称为题目。俗语说：“看人先看脸，看脸先看眼。”人们阅读文章，第一眼是看标题，加之标题常常是文章中心内容、基本思想的集中体现，因而标题也就成了文章的“眼睛”，在文章的结构中占有重要的地位。

在统计分析报告写作中，有相当部分作者不重视标题，这方面的通病有以下三点：一是标题无变化，格式老一套。例如：《关于×××的分析》或《关于×××的调查》，这类标题大家可用，年年可用，这种公文式的标题，显得十分呆板。二是题文不一致。往往是题意过宽或题意过窄。三是缺乏吸引力。由于标题无变化、格式陈旧，对读者没有吸引力。

制作好标题，一般来说有正题和辅题。正题(也叫主题或大标题)，辅题包括引题和副

题。引题(也叫肩题、眉题或小题)是正题的引子;副题(也叫次题或提要题)是正标题的辅助标题,用于进一步补充和说明正题,使正题的意思更完整。要使标题新颖醒目,扣人心弦,增加吸引力,引起人们的重视,可以采取以下一些方法。

(1) 多用“论点题”和“事实题”,少用“对象题”。试比较以下两题(乡镇工业大有作为;关于乡镇工业的调查)。

(2) 适当采用“设问题”。试比较以下两题(商品库存为什么升高;商品库存情况的分析)。

(3) 用具体事实做标题。试比较以下两题(我县夏粮增长四千万斤;我县夏粮获得丰收)。

(4) 用突出的事实做标题。试比较以下两题(我区工业总产值突破一千亿大关;我区工业生产大幅度增长)。

(5) 加重语气。试比较以下两题(我市蔬菜价格猛涨 26%;我市蔬菜价格上涨 26%)。

(6) 运用对比手法。试比较以下两题(改革前长期亏损,改革后一年盈利上百万;改革后我厂扭亏为盈全年盈利过百万)。

(7) 适当运用比喻。试比较以下两题(××地区大力营造“绿色宝库”;××地区开展植树造林情况)。

(8) 适当运用诗词、成语、古语、警句。试比较下面两题(安得广厦千万间,黎民百姓尽开颜;我区房地产情况调查)。

(9) 适当运用副题。试比较以下两题(××县葡萄生产情况——今年全县葡萄产量可达 25 万吨,比上年增长 24%;××县葡萄生产情况)。

(10) 适当运用提示语和有强调作用的语句。试比较以下两对标题(请注意:我区耕地面积大量减少;我区耕地面积大量减少)。

3. 开头

人们常说:写文章是“头难起,尾难收”。统计分析报告的开头,有如下三种方法:

(1) 开门见山,直叙主题,使之看了开头的话就能领悟到题旨。这是最常用的一种开头方法,具体的使用方式有:

① 起笔点题。一开始就点出基本事实。

② 亮出观点。开门见山,提出一个大家关心的问题,引出文章的主要内容和基本观点。

③ 强调意义。通过议论说明事件的重要性、突出该文章的中心内容、作用和意义。

④ 总说全文。这种开头,把全文所要阐述的内容做概括的介绍。使读者在开始即能了解总的情况,也为全文的论述定下基本的格局。

(2) 造成悬念。先提出一个问题,为什么会这样?引起读者的注意和思考。

(3) 交代动机。如“为了……”。

4. 正文

正文是报告的中心。用几个论点来阐述主题,每个大的论点有哪些小论点,大小论点如何连贯,问题如何排列,等等。统计分析报告一般由四个部分构成。

(1) 现状(问题、矛盾、难点)。

(2) 原因。

（3）对国民经济的影响。

（4）政策性建议。

5. 结束语

统计分析报告的结束语，大致有以下几种写法。

（1）总结全文，深化主题。

（2）表明态度，提出建议。

（3）展望前景，提出看法。

（4）强调问题，引起重视。

（5）水到渠成，得出结论。

（6）呼应开头，首尾圆合。

（7）事物未来，做出预测。

四、统计分析报告的写作要求

统计分析报告的写作要求有以下几点。

（一）主题要突出

主题是统计分析报告的中心思想或基本论点，它像一根红线贯穿全文，是文章的灵魂与统帅。统计分析报告要根据统计研究的任务，抓住要解决的主要矛盾及矛盾的主要方面，开展分析工作。内容要紧扣主题，从统计资料反映的复杂社会经济现象中，抓住重点问题，突出主题思想加以阐述。

（二）材料和观点要统一

统计分析报告必须以统计资料为依据，但不能搞资料堆砌，要用统计资料来说明观点。这就要求编写统计分析报告必须处理好材料与观点的关系。统计资料要支持报告所说明的观点，而观点要依据统计资料，做到材料与观点的辩证统一。如果材料与观点脱节，便失去了统计分析报告的说服力。

（三）判断推理要符合逻辑

统计分析报告的准确性，不仅是运用的统计数字要准确可靠，而且要准确地说明社会经济现象的本质和发展变化的规律。这就要求编写统计分析报告要在统计资料的基础上进行深入分析，运用推理和判断的逻辑方法。判断是以准确的统计数字为依据的；推理是以充分的依据为前提的。正确的判断和推理，从事物发展上说，就是要有根有据，符合客观的规律性；从思维发展上说，就是要实事求是，合乎事物的逻辑性。判断和推理的结果，前后不能矛盾，左右不能脱节，要如实反映客观事物的内在联系。

（四）结构要严谨

结构要严谨，是指统计分析报告内容的组织、构造精当细密，无懈可击，甚至达到“匠心经营，天衣无缝”的地步。这就要求首先要思想周密，没有“挂一漏万”“顾此失彼”；其次

要组织严谨,没有“颠三倒四”“破绽百出”。因此,结构能否严谨,首先取决于作者思想认识和思路是否清晰、严密。作者只有充分认识与掌握事物发展的内在规律,才能把它顺理成章地表达出来。

(五) 语言要生动、简练

统计分析报告的质量高低,首先在于内容正确,其次还要讲究词章问题。如果用词烦琐、语言不通、词不达意,就不能较好地表述分析的结果。所以,写一篇较好的分析报告,要善于用典型的事例、确凿的数据、简练的辞藻、生动的语言来说明问题。切忌文字游戏、词句堆砌,形式排比、华而不实。

(六) 报告要反复研究、修改

写统计分析报告与其他文章一样,必须反复研究和反复修改,做到用词恰当,符合实际。统计分析报告要进行反复研究和修改的目的,是检查观点是否符合政策,材料是否真实可靠,文章结构是否严密,文字是否言简意明,表达是否准确得当。只有反复修改,才能写出好的统计分析报告。

第三节　统计分析报告的举例

实例一

我国统计学类专业本科教育现状的调查与分析

教育部统计学类专业教学指导委员会

一、调查的背景与基本概况

近年来,随着统计学上升为一级学科以及“大数据”时代的到来,统计学的重要性和发展前景日益得到社会承认。与前几年相比,统计学科点和招生数均呈现出“大跃进”的发展态势。

根据有关统计资料,2010 年设置统计学本科专业的学校(包括理学类和经济学类)共有 172 个。其中,授经济学位的大约 120 个,授理学学位的大约 70 个(两者合计超过 172 个,这是因为其中有些学校既有经济学位的授权点,又有理学学位的授权点)。每年大约有8 000 名统计学专业的毕业生。

2012 年教育部对原有的本科专业目录进行修订,在最新的 2012 版本科专业目录中,冠有“统计学”名称的专业有以下 3 个:一是属于经济学门类经济学类专业下的“经济统计学”专业(代码 020102),另外两个则是属于理学门类统计学类专业下的“统计学专业”(代码 071201)与“应用统计学专业”(代码 071202)。新版专业目录制定后,2012 年,授理学学位的统计学专业点和应用统计学专业点分别猛增到 194 个和 158 个,授经济学位的经济统计学专业点也有 164 个。2013 年又有 26 个新的统计学学科点通过教育部的评

审。因此，目前已通过的全国统计学类专业本科点共有542个。上述专业点若全部招生，学生数将比2010年翻3番。

在统计学科点迅速增加、统计本科生招生规模迅速扩大的背景下，如何保证统计教育的质量成为迫切需要研究解决的问题。作为提高教育质量的一项重要举措，教育部提出要在本年内制定高等学校本科各专业的国家教学质量标准，作为今后高等学校专业准入、学科建设与教学质量评价的基本标准。教育部委托教育部各专业教学指导委员会完成此项工作，并将其列为本年度教育部与本届各专业教学指导委员会工作的重中之重。

为了使统计学专业教学质量的国家标准，既反映国际的先进水平满足经济社会发展的需要，又适合我国的国情，在制定教学质量标准时不能“闭门造车”，而是非常有必要充分了解全国各高校统计学本科教育的现状。为此，教育部统计学类专业教学指导委员会组织了一次全国统计学类专业的教学情况调查。此次调查由厦门大学数据挖掘研究中心具体实施，调查得到了全国各统计学办学单位的大力支持。

本次调查的问卷分为教学单位问卷与学生问卷。单位问卷主要调查各校统计学专业的基本状况，包括师资、学生人数和实验室设备等，以及开设统计学核心课程的情况。学生问卷主要调查学生对统计学核心课程的意见，以及毕业后的求职或就业的去向。

单位问卷向324所学校的统计学科点负责人发放[①]，最终回收了87所高校、116个专业点的有效问卷。教学单位的有效问卷几乎涵盖了所有教学指导委员会委员所在的高校以及统计学办学历史较久的高校，包括了不同层次学校与不同类型的统计学专业。因此，可以认为，通过此次调查了解的教学单位情况具有较强的代表性。

学生问卷委托各单位协助发放，原则上每个单位10份，其中本科高年级学生5份、近3年毕业的本科生(包括就读于本校的在校研究生)5份。最终收回学生问卷900多份(涵盖86所高校)。需要说明的是，其中有一些教学单位将问卷发放给了毕业班全体同学填写。为保证问卷的代表性，我们在这几所高校寄回的学生问卷中，随机抽取了10份。最后用于分析的学生有效问卷为654份。学生问卷中在校生较多，已毕业的学生较少，毕业的学生中在校攻读研究生的居多。另外，按不同层次高校和不同类型专业分组后，各细分类的有效问卷偏少。因此，学生问卷调查结果的代表性可能存在一些偏差，但仍有一定的参考价值。现将有关调查结果反馈如下，供大家参考。

二、学校与专业分布情况分析

表10-1与表10-2分别列出了本次调查中，统计学重点学科与不同的统计学专业在不同层次高校的分布情况。从表中可以看出，目前我国统计学的国家级重点学科点主要分布在“985”与“211”高校。二、三本院校的统计学专业点与一本以上院校的统计学专业点各占一半左右。二、三本院校的应用统计学专业点明显多于一本以上院校，统计学专业和经济统计学专业的分布则无明显差别。本次调查结果表明：在我国统计学类本科专业教育中，二、三本层次的院校占有相当大的比重，在制定统计学类专业教学质量国家标准时有必要充分考虑这一层次院校的实际情况。

---

① 有的学校同时有多个统计学类学科点。

表 10-1　统计学重点学科在不同层次院校分布情况

学校类别	国家一级重点学科	国家二级重点学科	省部级重点学科	其他	总计
二、三本院校	0	0	11	30	41
一本院校	0	1	11	6	18
“211”高校	1	1	3	8	13
“985”高校	4	4	5	2	15
总计	5	6	30	46	87

注：① 统计学的国家一级重点学科目前尚未评定，此处的国家一级重点学科包括“应用经济学”和“数学”一级重点学科；研究生目录修订之前，概率统计包括在数学学科下，经济统计包括在应用经济学下。

② 统计学的国家二级重点学科包括应用经济学下的“统计学”和数学下的“概率统计学”重点学科。

③ 表中的学校分类“一本院校”不包括属于“211”和“985”的一本院校；“211”高校不包括既是“985”又是“211”的高校。以下相同，不一一标注。

表 10-2　不同类型统计学专业点分布情况

学校类别	经济统计学	统计学	应用统计学	总计
二、三本院校	20	23	15	58
一本院校	10	9	4	23
“211”高校	7	7	2	16
“985”高校	2	15	2	19
总计	39	54	23	116

三、专业教师与专业实验室配备情况分析

专业教师与教学实验室的配备情况是国家教学质量标准中考察新专业准入的基本硬件。表 10-3 和表 10-4 列出了本次调查各单位专业教师配备情况和平均生师比。从表中可以看出，我国统计学类专业教师的配备有以下两个特点。

表 10-3　专业教师配备情况

学校类别	平均教授数	平均副教授数	平均讲师数	平均教师总数	有效样本个数
二、三本院校	2.75	5.17	6.38	14.91	53
经济统计学	3.17	5.28	6.67	15.78	18
统计学	2.75	4.95	6.05	14.50	20
应用统计学	2.27	5.33	6.47	14.40	15
一本院校	5.30	7.32	6.78	20.00	23
经济统计学	6.00	7.56	7.30	20.70	10
统计学	5.22	6.89	6.33	19.33	9
应用统计学	3.75	7.75	6.50	19.75	4
“211”高校	4.50	8.13	5.44	18.56	16
经济统计学	4.29	7.43	3.86	16.00	7
统计学	4.00	7.29	6.00	18.00	7
应用统计学	7.00	13.50	9.00	29.50	2
“985”高校	6.79	7.89	6.67	21.21	19
经济统计学	12.00	13.00	12.00	37.00	2
统计学	5.53	6.67	5.93	17.93	15
应用统计学	11.00	12.00	6.50	30.00	2
总计	4.23	6.50	6.37	17.57	111

表 10-4　平均生师比情况

学校类别	生　师　比	有效样本个数
二、三本院校	46.70	52
经济统计学	45.10	18
统计学	45.43	19
应用统计学	50.22	15
一本院校	28.29	23
经济统计学	33.33	10
统计学	23.76	9
应用统计学	25.89	4
“211”高校	17.47	16
经济统计学	16.27	7
统计学	20.83	7
应用统计学	9.93	2
“985”高校	15.94	17
经济统计学	10.42	1
统计学	17.09	14
应用统计学	10.67	2
总计	33.61	108

注：表中的生师比按以下公式估算 生师比＝(2013 年本科招生数×4)/平均专业教师数。

(1) 随着学校层次的提高，平均教师总人数与高级职称教师的人数均有提高的倾向，而生师比则有降低的倾向。这表明，层次越高的学校师资水平越高，师资配备情况也越好。但需要指出的是，本次调查未对各单位的在学研究生人数进行调查，这里的生师比是按照调查结果估算的统计学专业本科生与统计学专业教师的比例。事实上，因为层次越高的学校博士生和硕士生越多，所以高层次学校按照一定比例折算后的包括研究生在内的生师比会远远高于表中所列的数字。因此，在今后的调查中，应当注意增加对研究生在校人数的调查，以弥补此次调查的不足。

(2) 一般而言，经济统计学专业配备的平均教师较多，应用统计学专业的教师数次之，统计学教师数则相对较少。之所以出现这种情况，与经济统计学和应用统计学专业除了需要一般统计理论与方法课程的教师外，还需要相关专业的教师有关。另外，此次调查对专业教师的定义尚不够明确。经济类院校经济统计学专业所在的院校多为财经类院校，在这一类院校中承担非统计学专业统计学课程的教师可能比较多。

从本次调查各单位教学实验室情况看，除 1 个单位外，其他专业点均配备了实验室，而且绝大多数专业点(92 个单位)还配备了专属的实验室统一安装统计软件。从计算机数量看，除“211”高校平均每个专业点拥有的计算机数量较多外，其他层次的院校并无明显差异。从统计软件看，使用最多的是 SPSS，经济统计专业使用 Eviews 也比较普遍。使用 R 语言的单位主要是层次较高的院校和统计学专业。

四、统计学核心课程体系与教学内容分析

一个专业的核心课程体系与教学内容是该专业与其他专业区分的最主要标志，科学的专业核心课程体系对于指导各专业建设和评估各单位的办学质量有非常重要的意义。因此，关于这方面的调查是本次调查的核心内容。不同类型的统计学专业核心课程应有

所差别,在本次调查问卷中,经济统计学专业与统计学专业所列的课程是上一届统计学教学指导分委员会制定的两类统计学专业教学规范中规定的核心课程。应用统计专业是这次教育部本科目录修订后新增加的专业,其核心课程体系在参照统计学专业课程的基础上,适当增加了一些具有应用特点的统计课程。

从核心课程的开课情况看,经济统计专业的大多数单位能够按教学规范建议的课程体系开课,在回答问卷的 39 个经济统计学专业点中,统计学方法类的课程(包括:统计学导论、数理统计学、应用多元统计分析、应用时间序列分析、应用抽样技术和计量经济学课程)的开设单位均在 32 个以上。应用类经济统计课程的开设单位则相对较少,除开设国民经济统计学的专业点有 29 个外,其他课程的开设单位均在 20 个左右。统计学专业回答问卷的有 54 个单位,其中开设较多的课程为数理统计学、多元统计分析、随机过程、抽样技术、时间序列分析、统计计算与软件。应用统计学专业开设的课程介于经济统计学专业与统计学专业之间,稍感意外的是开设应用统计专题的单位较少,在回答问卷的 23 个单位中,只有 5 个单位开设了应用统计专题。这表明到目前为止,该专业的特色尚不够鲜明。

从未开足教学规范建议课程的主要原因看,认为"没有必要"的有 32 个单位,"师范力量不足"的有 32 个单位,"缺乏合适教材"的有 15 个单位。

高等学校的办学要"以学生为本",为了进一步完善与充实课程体系与教学内容,有必要充分倾听学生的意见。为此,我们还就学生对目前统计学专业所开设课程的满意程度进行了调查,并请学生对所修过的课程的有用性做出评价。

从对学生开设课程总体满意度调查汇总的结果可以看出,对目前的课程体系,认为非常满意和比较满意的学生所占的比重为 78.7%。一般来讲,层次越高的学校学生满意度也越高。就不同类型的专业看,二、三本层次和"985"层次经济统计学专业学生满意度较高,尤其是"985"高校满意度达 100%。"211"和一本层次统计学专业学生满意度较高。除"985"高校外,其他各层次应用统计学专业学生的满意度均比较低。

在各门课程中最受欢迎的是"统计计算与软件",82%修过该课程的同学认为课程很有帮助。以下依次是"多元统计分析""概率论与数理统计""抽样技术""时间序列分析""计量经济学""会计学""统计学导论(原理)""宏微观经济学"等。比较不受欢迎的则有"测度论""复变函数"和"解析几何",认为这些课对自己有所帮助的学生所占的比例在 40%以下。

就业意向不同或者已经参加工作与否,对各门课程的认可程度也可能有差别。图 10-1 和图 10-2 所示为不同类型学生对各门课程的认可程度。

从图 10-1 可以看出,毕业生与在校生对于课程的评价基本相似,但对多数课程的评价在校生略高于毕业生。统计学导论和数理统计是少有的毕业生评价高于在校生的两门课程。

从图 10-2 可以看出,不同就业意向的学生对具体课程的认可程度各有特点。准备继续深造的学生对"高等数学"和"统计学导论"等课程的认可程度明显高于其他就业意向的学生。就业意向为企业的学生对"企业经营统计学"课程的认可程度明显高于其他就业意向的学生。就业意向为金融机构的学生对各门课程的认可程度曲线相当平滑,其他就业意向的学生对各门课程的认可程度曲线则围绕该曲线上下起伏。另外,从问卷调查的结果还可知,经济统计专业与应用统计专业的学生对数学类基础课程的评价相对较低,对经济学和会计学等课程的评价则相对较高。

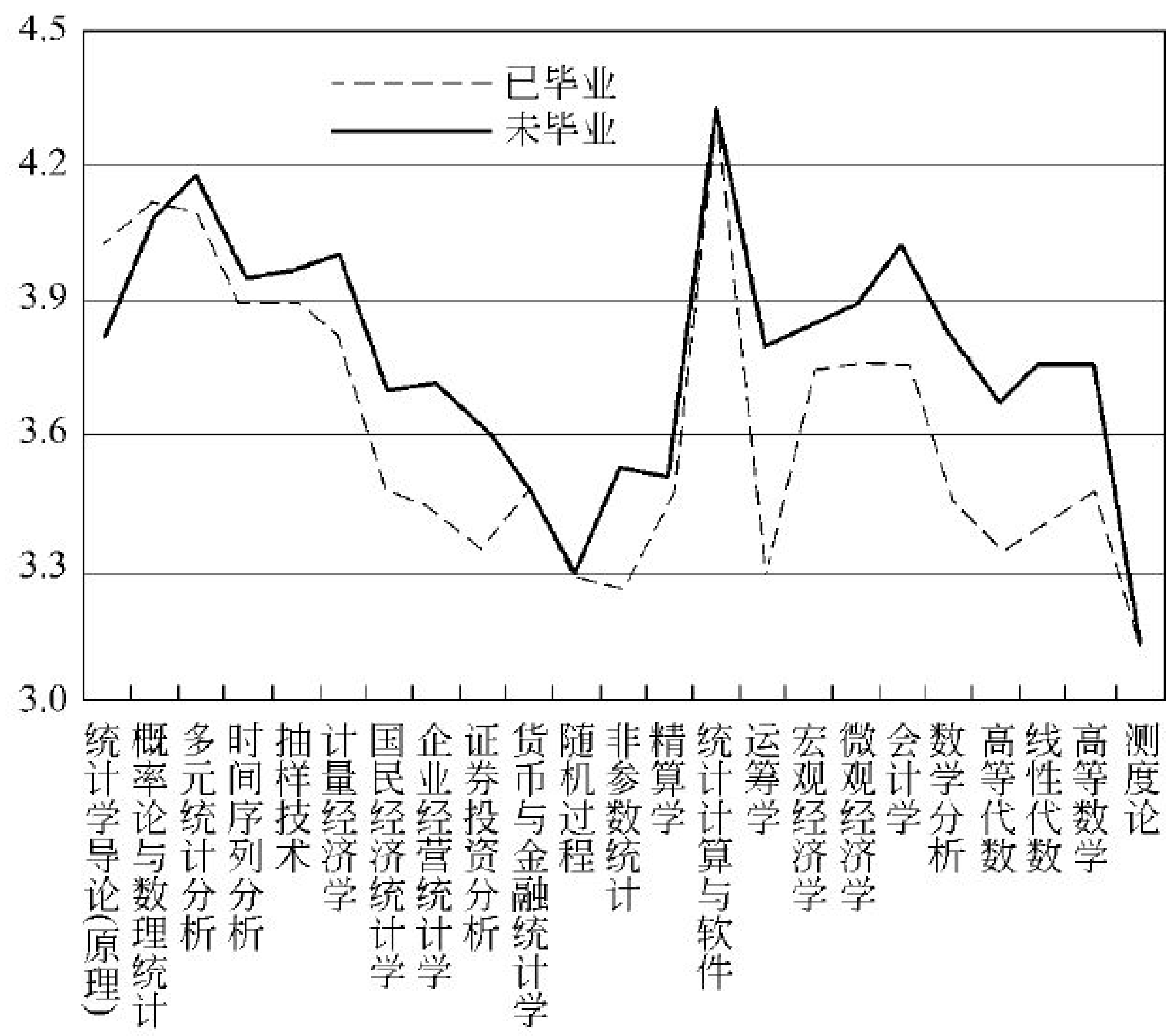


图 10-1　毕业生与在校生对各门课程的认可程度

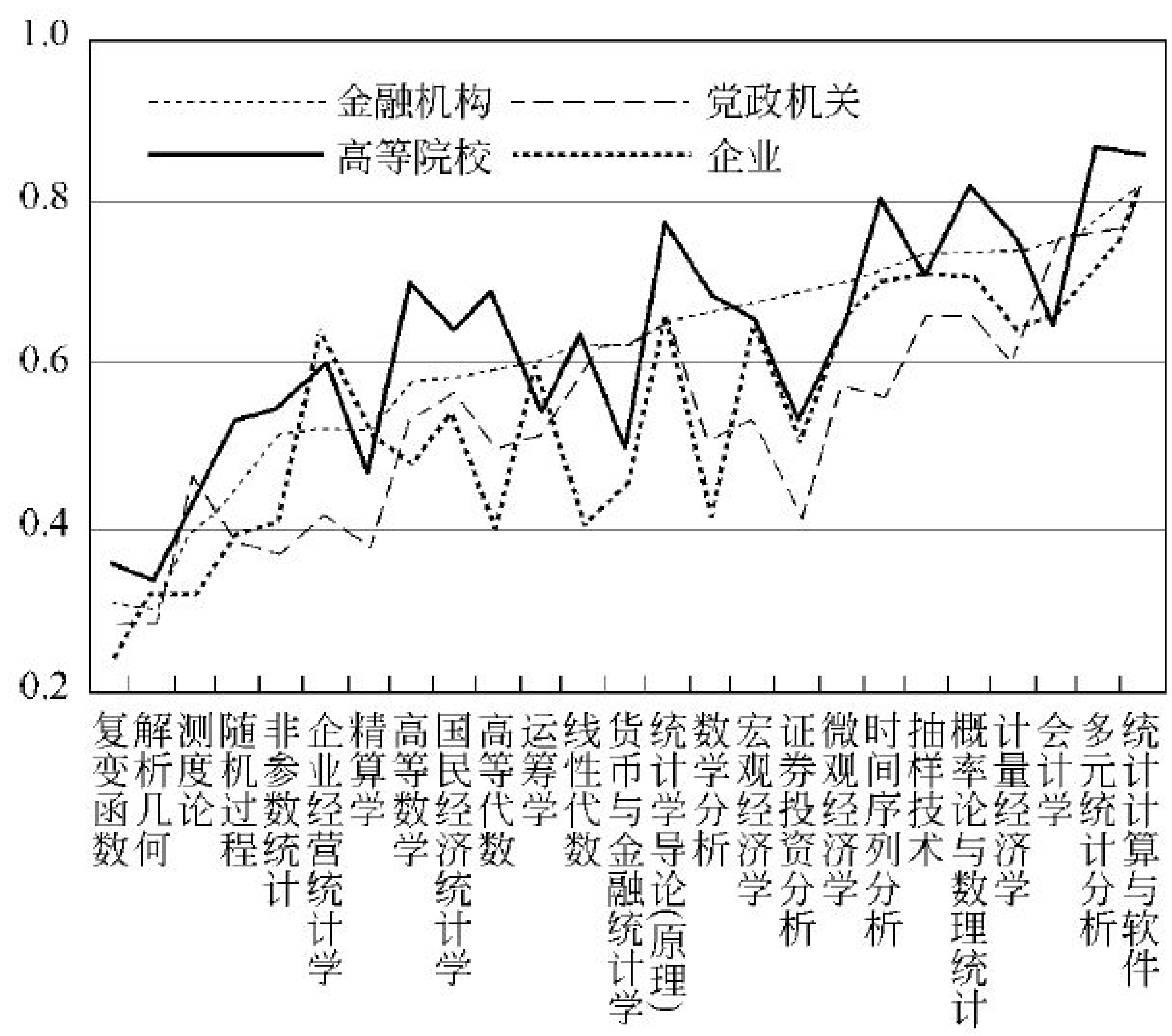


图 10-2　不同就业意向的学生对课程的认可度

五、招生与就业情况分析

高等学校最主要的任务就是培养人才。二、三本院校与普通一本院校统计学专业点的本科平均招生数明显高于“985”高校与“211”高校。这是因为，高层次的学校招收了较多的硕士生与博士生。

我国目前统计学类专业总体就业情况还是比较好的，总的平均一次就业率(含升学)为 87.84%，属于各类专业的中上水平。但不同层次的学校有所差别，一般来讲，层次越高的学校一次就业率也越高。从专业看，多数院校的经济统计学专业和应用统计学专业

的一次就业率高于统计学专业。

从 2013 年我国统计学专业本科毕业生的主要去向看,依次为金融机构(占 37.61%)、其他企业(占 25.99%)、高校(占 27.39%)和党政机关(占 10.7%)。从不同层次的高校看,层次较低的学校的统计与专业毕业生去企业的较多。随着学校层次的提高,到高等学校继续深造的毕业生所占的比重也在提高。

初次就业的薪酬可以在一定程度上反映社会对毕业生的认可程度。这次我们也对此做了调查,如表 10-5 所示。但由于分组的有效样本偏少,可能缺乏足够的代表性,仅供参考。目前统计学类专业本科毕业生初次就业的平均月薪为 3 333 元。一般而言,随着学校层次提升,初次就业的薪酬水平也相应提高。另外,较高层次学校经济统计学专业毕业生的平均月工资高于其他专业。二、三本院校则呈现相反情况。

表 10-5 统计学专业毕业生初次就业的薪酬情况

学校类别	平均月工资/元	有效样本数	学校类别	平均月工资/元	有效样本数
二、三本院校	2 975	96	经济统计学	4 483	6
经济统计学	2 858	26	统计学	3 594	18
统计学	2 989	64	应用统计学	3 000	1
应用统计学	3 333	6	985 高校	4 806	17
一本院校	3 193	14	经济统计学	6 000	4
经济统计学	3 483	6	统计学	4 650	12
统计学	2 560	5	应用统计学	1 900	1
应用统计学	3 667	3	总计	3 333	152
“211”高校	3 784	25			

(资料来源:《统计研究》2015 年第 2 期)

实例二

中华人民共和国 2015 年国民经济和社会发展统计公报(节选)

中华人民共和国国家统计局

2016 年 2 月 29 日

2015 年,面对错综复杂的国际形势和艰巨繁重的国内改革发展稳定任务,党中央、国务院团结带领全国各族人民,按照“五位一体”总体布局和“四个全面”战略布局的总要求,牢固树立和贯彻落实创新、协调、绿色、开放、共享的发展理念,适应经济发展新常态,坚持改革开放,坚持稳中求进工作总基调,坚持稳增长、调结构、惠民生、防风险,不断创新宏观调控思路与方式,深入推进结构性改革,扎实推动大众创业万众创新,努力促进经济保持中高速增长、迈向中高端水平,转型升级步伐加快,改革开放不断深化,民生事业持续进步,经济社会发展迈上新台阶,实现了“十二五”圆满收官,为“十三五”经济社会发展、决胜全面建成小康社会奠定了坚实基础。

初步核算，全年国内生产总值 676 708 亿元，比上年增长 6.9%。其中，第一产业增加值 60 863 亿元，增长 3.9%；第二产业增加值 274 278 亿元，增长 6.0%；第三产业增加值 341 567 亿元，增长 8.3%。第一产业增加值占国内生产总值的比重为 9.0%，第二产业增加比重为 40.5%，第三产业增加值比重为 50.5%，首次突破 50%。全年人均国内生产总值 49 351 元，比上年增长 6.3%。全年国民总收入 673 021 亿元。具体如图 10-3 和图 10-4 所示。

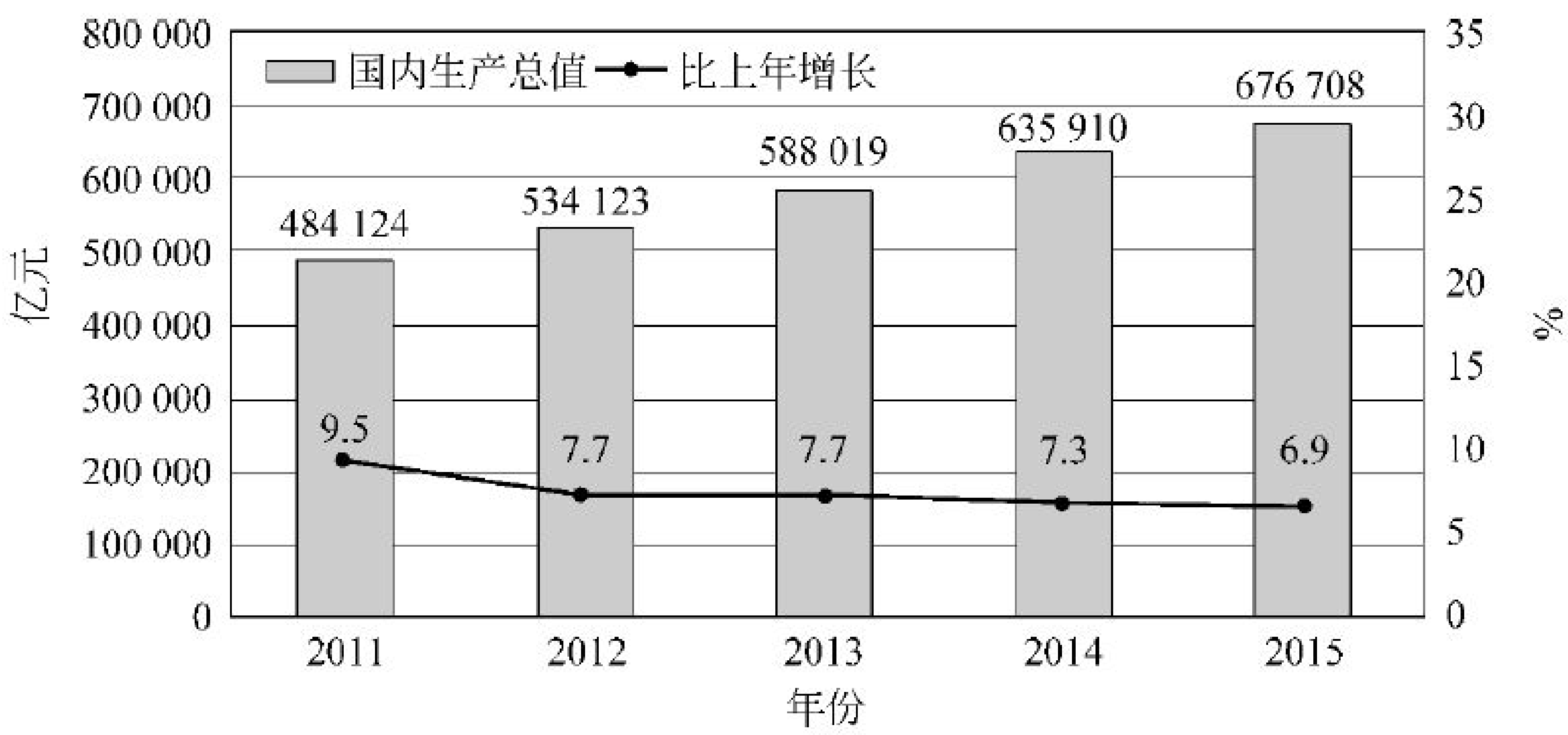


图 10-3 2011—2015 年国内生产总值及其增长速度

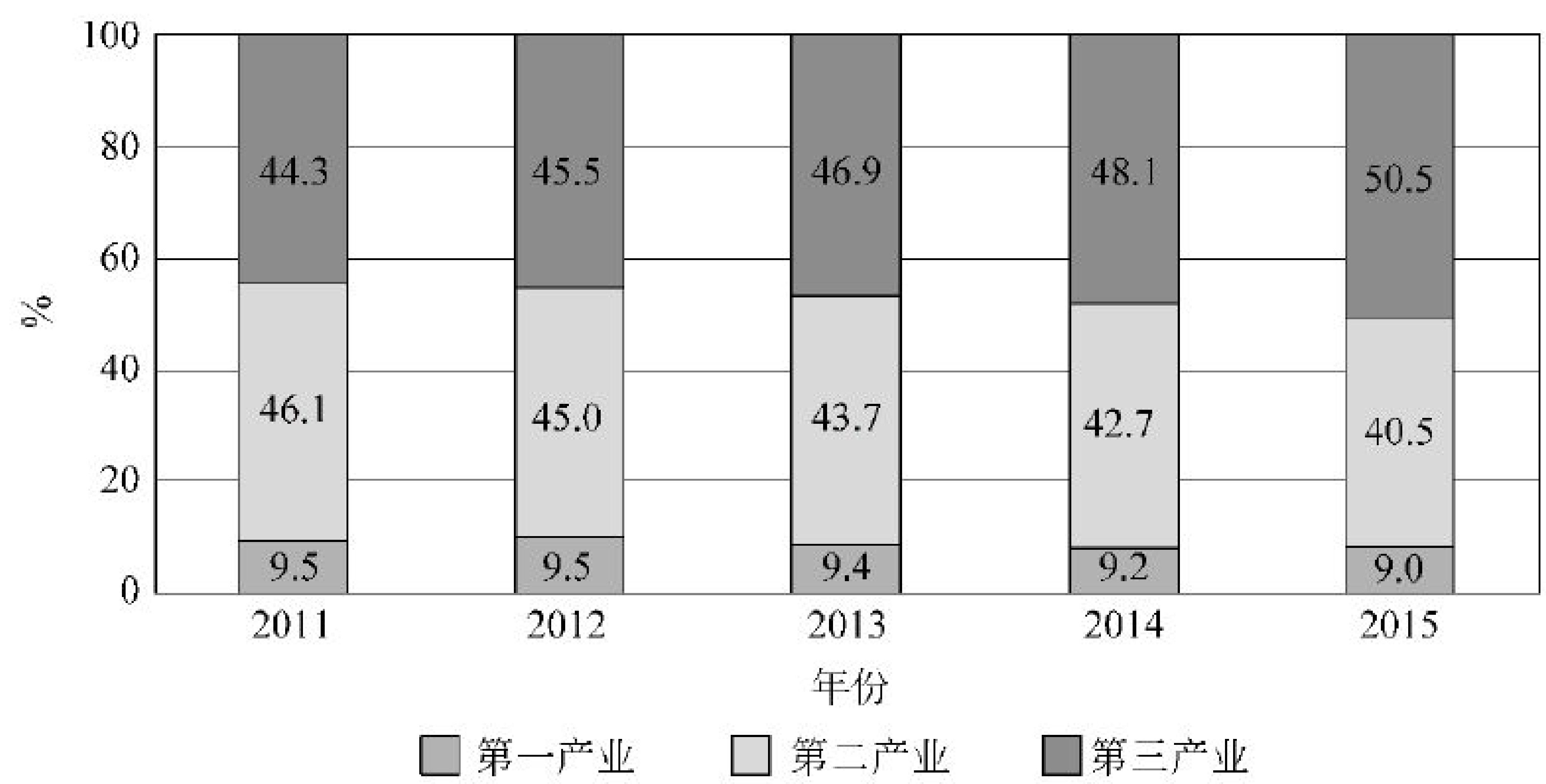


图 10-4 2011—2015 年三次产业增加值占国内生产总值比重

年末全国总人口 137 462 万人，比上年末增加 680 万人，其中城镇常住人口 77 116 万人，占总人口比重(常住人口城镇化率)为 56.10%，比上年年末提高 1.33 个百分点。全年出生人口 1 655 万人，出生率为 12.07‰；死亡人口 975 万人，死亡率为 7.11‰；自然增长率为 4.96‰。全国人户分离的人口 2.94 亿人，其中流动人口 2.47 亿人，人均预期寿命 76.34 岁。2015 年年末人口数及其构成如表 10-6 所示。

年末全国就业人口 77 451 万人，其中城镇就业人员 40 410 万人。全年城镇新增就业 1 312 万人。年末城镇登记失业率为 4.05%。全国农民工总量 27 747 万人，比上年增长 1.3%。其中，外出农民工 16 884 万人，增长 0.4%；本地农民工 10 863 万人，增长 2.7%。具体如图 10-5 所示。

表 10-6　2015 年年末人口数及其构成

全国总人口	年末数/万人	比重/%
全国总人口	137 462	100.0
其中：城镇	77 116	56.10
乡村	60 346	43.90
其中：男性	70 414	51.2
女性	67 048	48.8
其中：0～15 岁(含不满 16 周岁)	24 166	17.6
16～59 岁(含不满 60 周岁)	91 096	66.3
60 周岁及以上	22 200	16.1
其中：65 周岁及以上	14 386	10.5

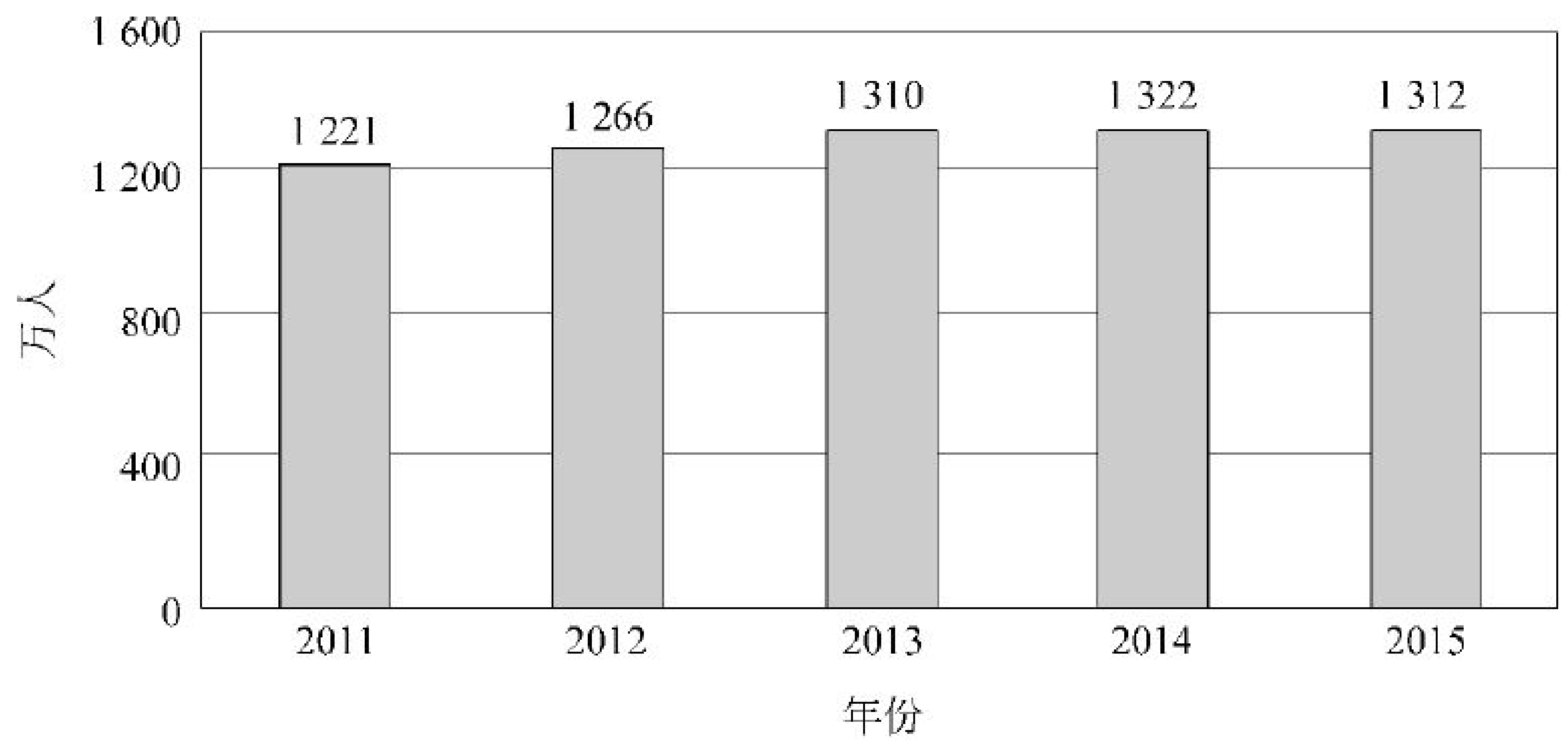


图 10-5　2011—2015 年城镇新增就业人数

全年全员劳动生产率为 76 978 元/人，比上年提高 6.6%。具体如图 10-6 所示。

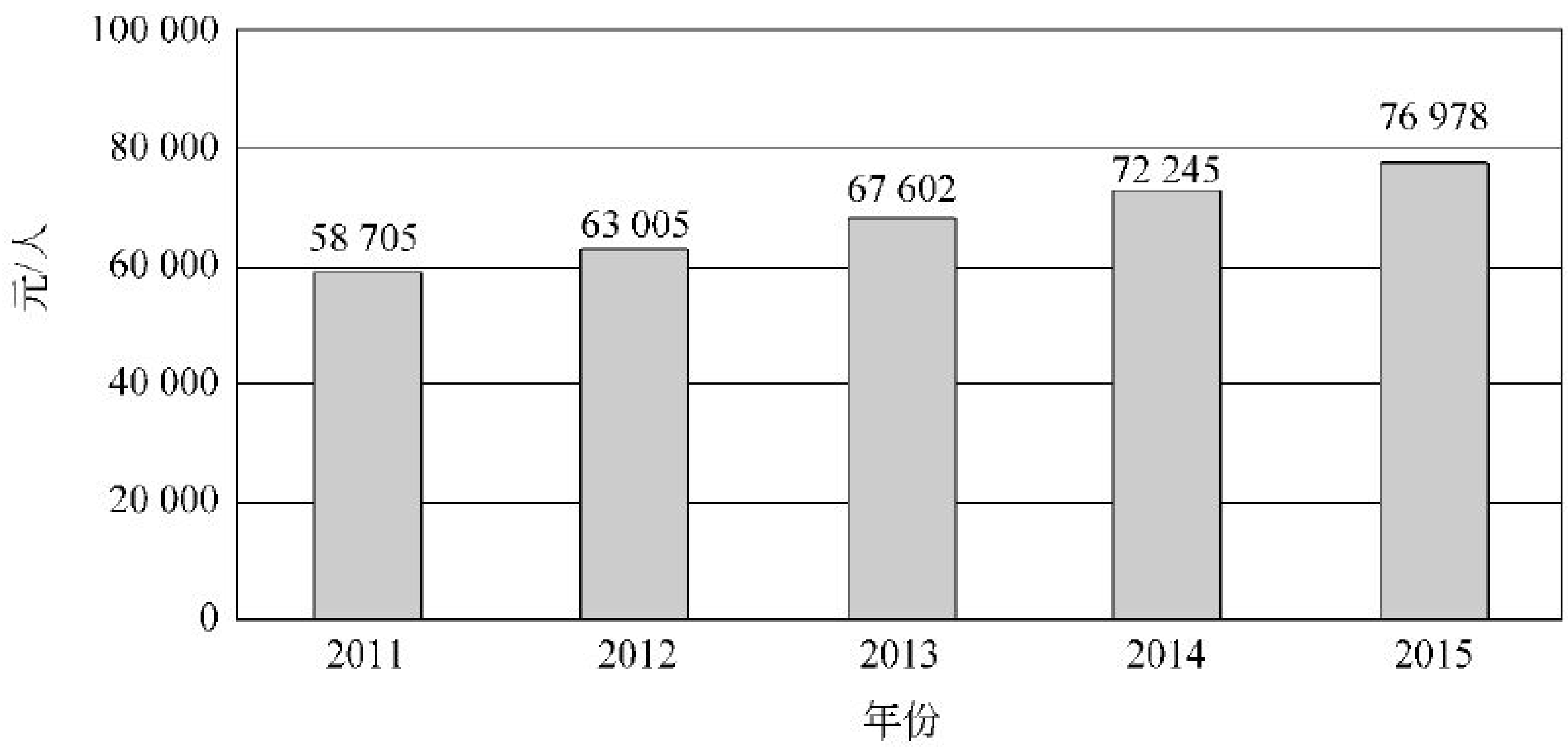


图 10-6　2011—2015 年全员劳动生产率

全年居民消费价格比上年上涨 1.4%，其中食品价格上涨 2.3%，固定资产投资价格下降 1.8%，工业生产者出厂价格下降 5.2%，工业生产者购进价格下降 6.1%，农产品生产者价格上涨 1.7%。具体如图 10-7 和表 10-7 所示。

表 10-7　2015 年居民消费价格比上年涨跌幅度　　%

指　　标	全　国	城　市	农　村
居民消费价格	1.4	1.5	1.3
其中：食品	2.3	2.3	2.4
烟酒及用品	2.1	2.0	2.3
衣着	2.7	2.8	2.3
家庭设备用品及维修服务	1.0	1.0	0.9
医疗保健和个人用品	2.0	1.9	2.3
交通和通信	－1.7	－1.6	－1.9
娱乐教育文化用品及服务	1.4	1.4	1.4
居住	0.7	1.0	－0.3

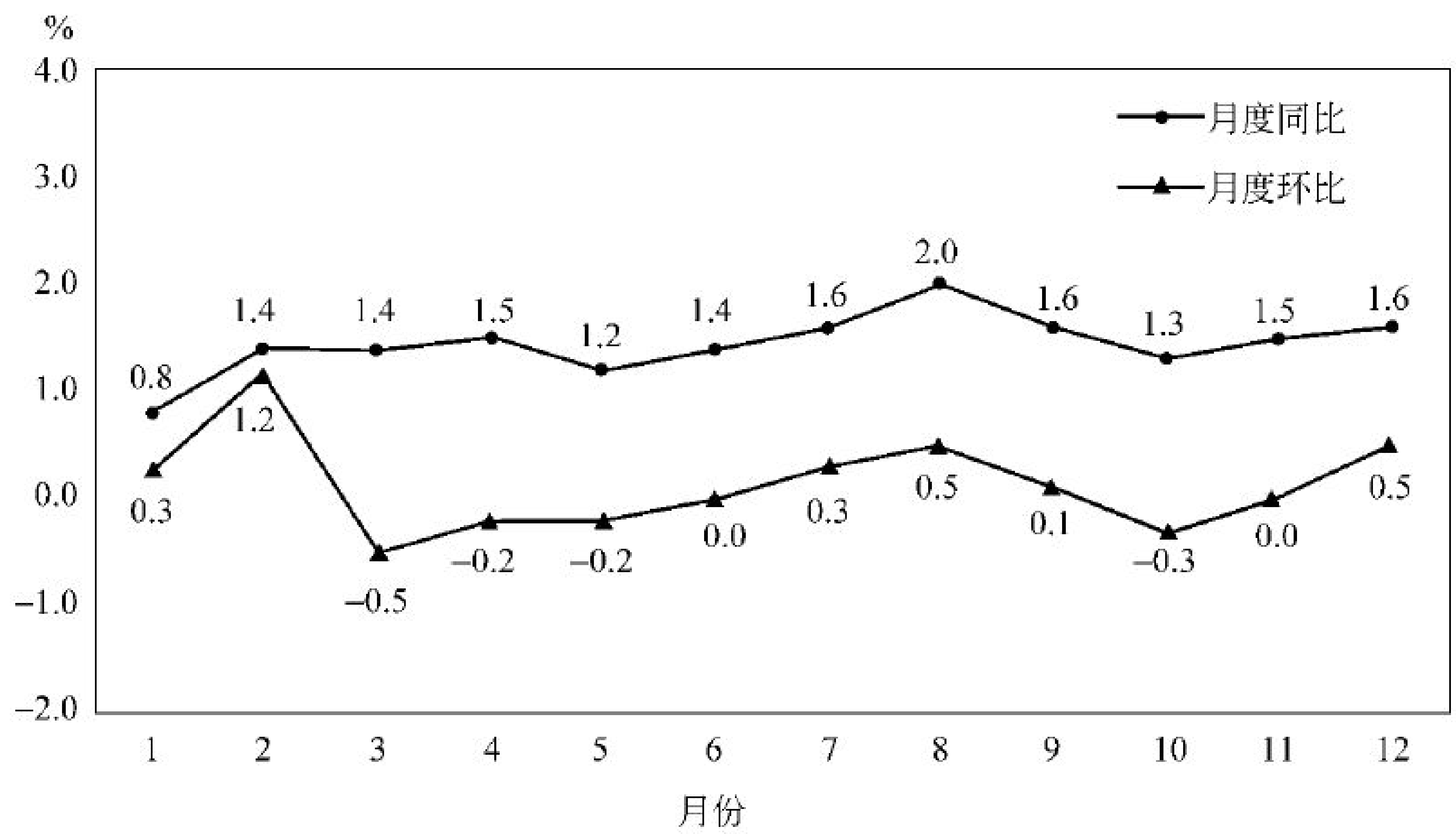
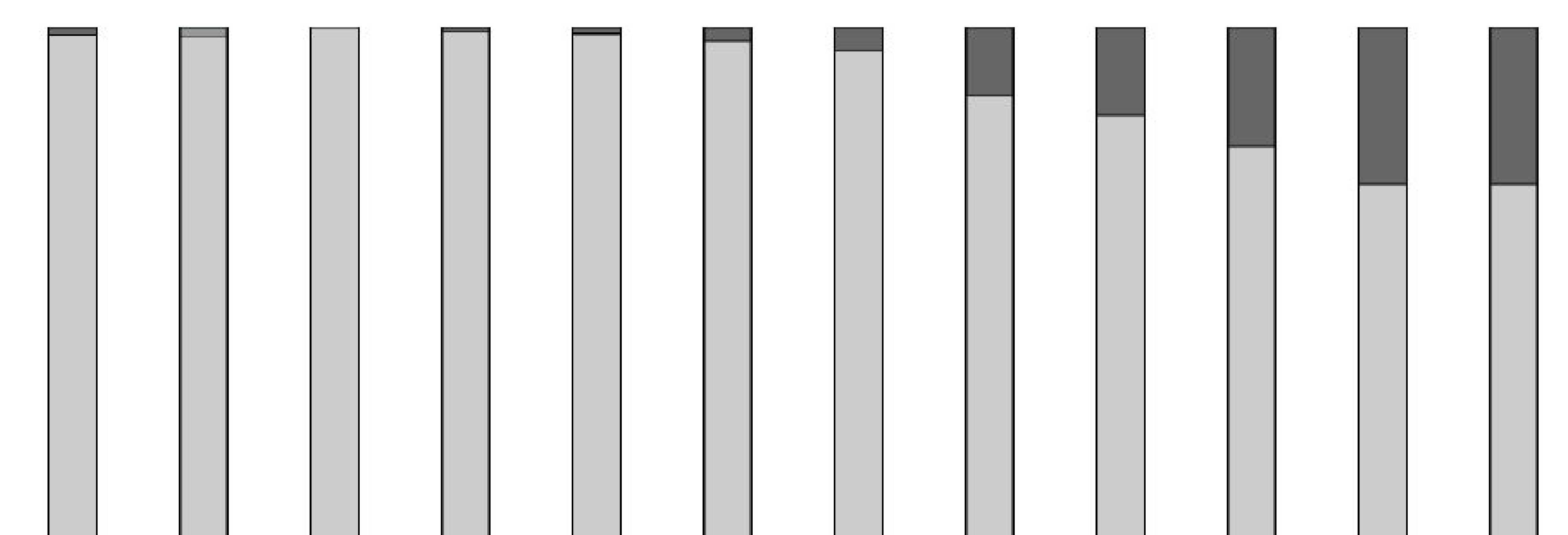


图 10-7　2015 年居民消费价格月度涨跌幅度

年末 70 个大中城市新建商品住宅销售价格月同比上涨的城市个数为 21 个，比年初增加 20 个；下降的为 49 个，减少 20 个，如图 10-8 所示。



	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
上涨	1	0	0	1	1	2	3	9	12	16	21	21
持平	0	1	0	0	0	0	0	0	0	0	0	0
下降	69	69	70	69	69	68	67	61	58	54	49	49

图 10-8　2015 年新建商品住宅月同比价格上涨、持平、下降城市个数变化情况

全年全国一般公共预算收入“152 217 亿元”,比上年同口径增加 8 324 亿元,增长 5.8%,其中税收收入 124 892 亿元,增加 5 717 亿元,增长 4.8%,如图 10-9 所示。

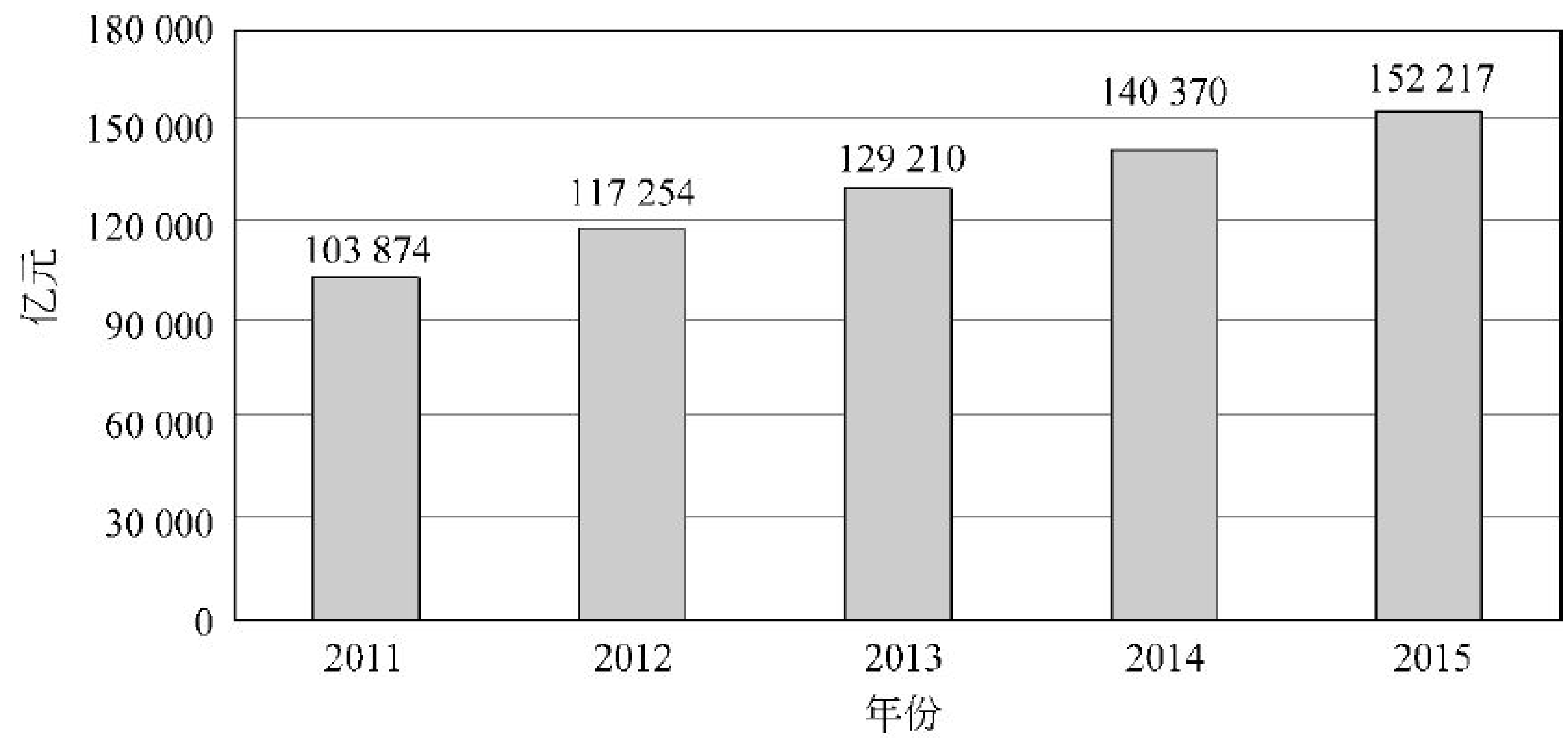


图 10-9 2011—2015 年全国一般公共预算收入

(注:图中 2011—2014 年数据为全国一般公共预算收入决算数,2015 年为执行数。)

年末国家外汇储备 33 304 亿美元,比上年年末减少 5 127 亿美元。全年人民币平均汇率为 1 美元兑 6.228 4 元人民币,比上年贬值 1.4%,如图 10-10 所示。

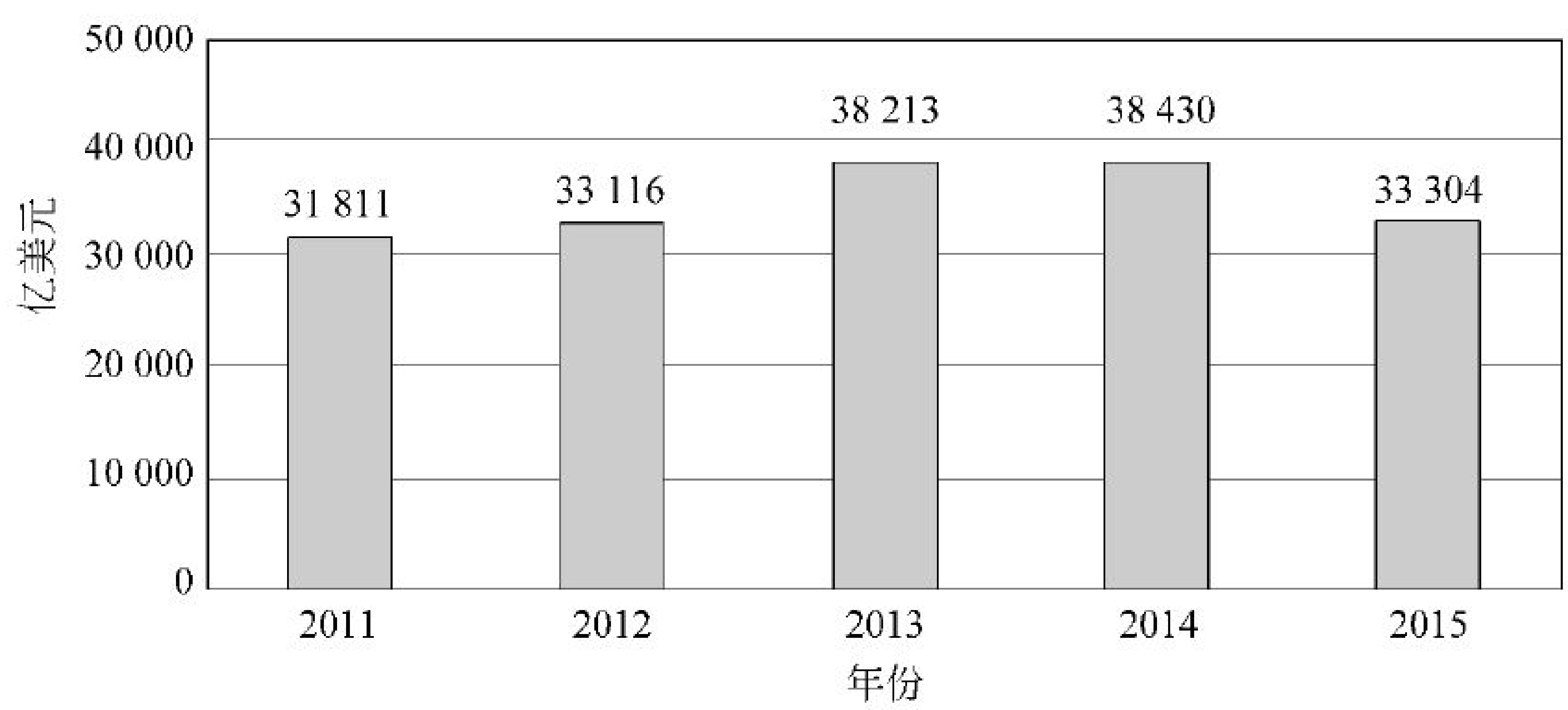


图 10-10 2011—2015 年年末国家外汇设备

(资料来源:国家统计局 http://www.stats.gov.cn/tjsj/zxfb/201602/t20160229_1323991.html.)

实例三

大学生生活费收支状况调查报告

第一部分 调查方案设计

一、调查方案

(1) 调查目的:通过了解大学生日常收入和消费的主要状况,为学校的助学政策提供参考,同时为大学生消费市场的开发提供一定的参考。

（2）调查对象：中国人民大学在校本科生。

（3）调查单位：抽取的样本学生。

（4）调查程序：

① 设计调查问卷，明确调查方向和内容。

② 分发调查问卷。随机抽取中国人民大学大一、大二、大三、大四在校本科生男、女各30人左右作为调查单位。

③ 根据回收有效问卷进行分析，具体内容如下：

a. 根据样本的生活费来源、分布状况的均值、方差等分布的数字特征，推断人大学生总体分布的相应参数。

b. 根据性别进行男女两个总体生活费均值之差的比较以及方差比的区间估计。

c. 根据大一、大二、大三、大四进行四个总体生活费均值之差及方差比的区间估计。

d. 绘制统计图形，使样本数据直观化并对统计量进行分析。

（5）调查时间：2013年4月20日—6月10日。

二、问卷设计

在经过我们共同的研究制定问卷雏形并征询老师的意见后，我们最终设计的问卷如下：

大学生收支调查问卷

××同学：

您好，请配合我们完成以下调查问卷，请在符合您的实际情况的选项下画"√"。

Q1. 您的性别：

A. 男　　B. 女

Q2. 您的年级：

A. 大一　　B. 大二　　C. 大三　　D. 大四

Q3. 您的月生活费支出在：

A. 300元以下　　B. 300～400元　　C. 400～500元　　D. 500～600元

E. 600～700元　　F. 700元以上

Q4. 您的生活费主要来源依次是：□→□→□→□

A. 父母　　B. 勤工俭学

C. 助学贷款　　D. 其他________（请注明）

Q5. 您的各项开支为(单位：元)：

A. 伙食费

B. 衣着

C. 书本资料及其他学习服品

D. 娱乐休闲

E. 日化服品(包括护肤、洗涤用品及其他日用小百货)

F. 其他

请排出您本学期支出的前三项：□→□→□

非常感谢您的合作！

三、问卷发放

本次调查我们采取分层抽样，对在校本科生各个年级男、女生各发放问卷30份左右；

我们在学一、学二、学八、学九共发放问卷 300 份,回收问卷 291 份,其中有效问卷共 265 份。现将各年级男女生回收有效问卷具体情况介绍如下:

大一: (男生)26 份 (女生)31 份

大二: (男生)34 份 (女生)40 份

大三: (男生)31 份 (女生)32 份

大四: (男生)41 份 (女生)30 份

总计: (男生)132 份 (女生)133 份

四、数据整理

为了便于用计算机进行数据处理,我们用数字代码来表示问卷信息,为了便于统一,对于问卷答案"A""B""C""D""E""F",我们分别用"1""2""3""4""5""6"表示(答案缺省项为空项)。例如:我们用"1"表示男性,用"2"表示女性;各个年级也分别用"1""2""3""4"来表示。

数据具体情况见附录。

第二部分 数据分析

根据以上整理的数据,我们进行数据分析。我们设样本一为抽样总体,样本二为男生的抽样总体,样本三为女生的抽样总体。

一、对生活费水平的分析

(一) 对样本一的分析

由整理后输入计算机的数据,我们绘制出样本一生活费水平的频数分布表(表 10-8)和直方图(图 10-11),结果如下。

表 10-8 样本一生活费水平的频数分布表

按支出分组/元	频数/人	频率累积/%
300 以下	4	1.51
300～400	41	16.98
400～500	74	44.91
500～700	62	68.30
600～700	33	80.75
700 以上	51	100.00
合计	265	100.00

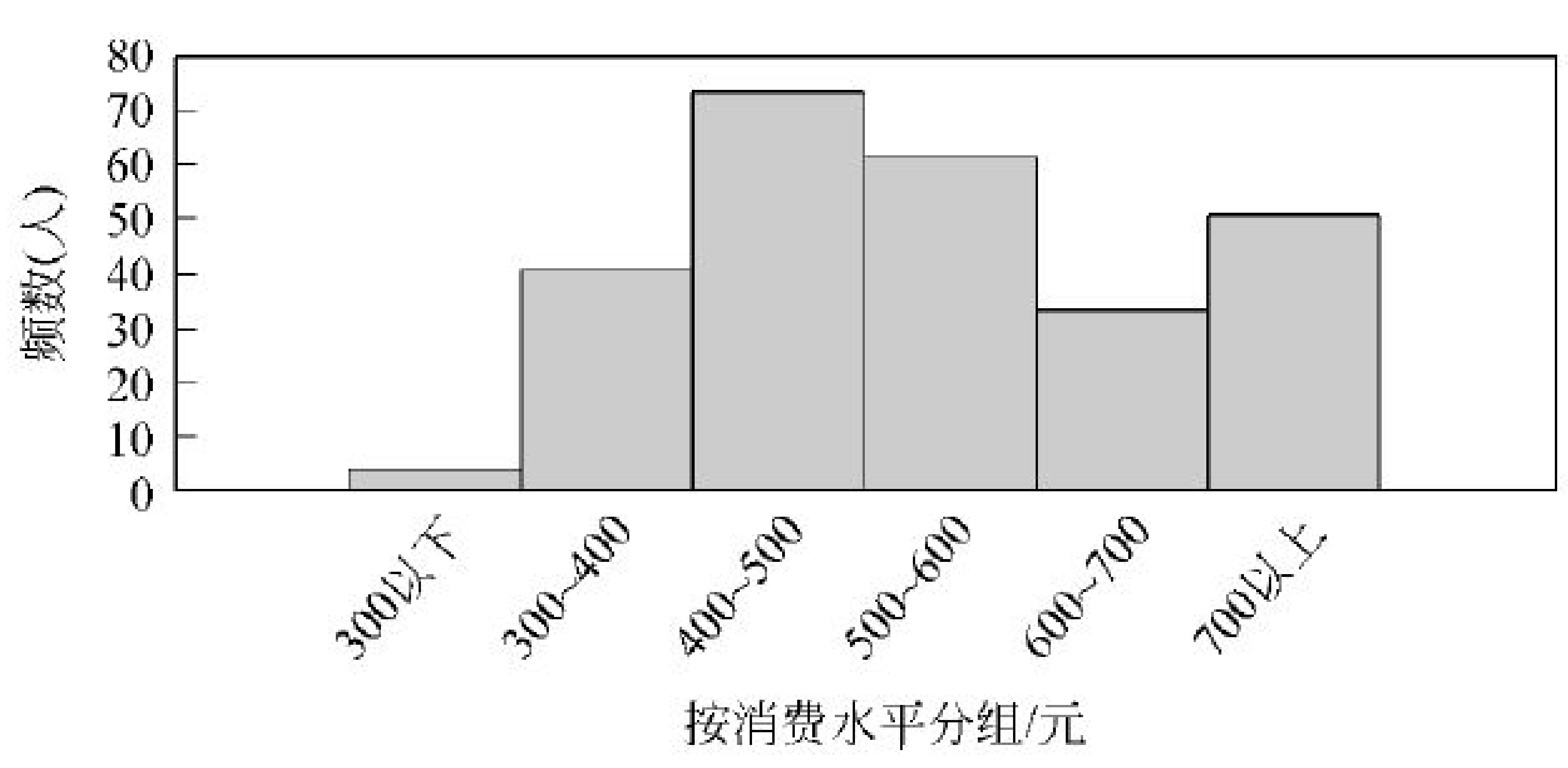


图 10-11 样本一消费水平的直方图

由图 10-11 可以看出：样本一（本科生抽样全体）月生活费 400～500 元所占频数最高。

我们进一步分析月生活费的平均水平，得出结果（表 10-9）。

表 10-9　样本一（总体）平均月生活费置信区间的构造表

月生活费水平/元	频数统计	组中值	描述统计输出结果	
300 以下	4	250	样本数据个数	265
300～400	41	350	样本标准差	138.57
400～500	74	450	样本均值	537.55
500～600	62	550	置信水平	95%
600～700	33	650	自由度	264
700 以上	51	750	t 值	1.968 992 365
			抽样平均误差	8.512 287 434
			误差范围	16.760 628 96
			置信下限	520.789 371
			置信上限	554.310 629

从上述分析可知：我们有 95% 的把握认为人大本科生的月生活费平均水平在 520.79～554.31 元。

（二）对样本二的分析

由整理后输入计算机的数据，我们绘制出样本二月生活费水平的频数分布表（表 10-10）和直方图（图 10-12），结果如下。

表 10-10　样本二月生活费水平的频数分布表

序号	月生活费水平/元	接收	频数/人	频率累积/%
1	300 以下	1	1	0.76
2	300～400	2	24	18.94
3	400～500	3	38	47.73
4	500～600	4	30	70.45
5	600～700	5	17	83.33
6	700 以上	6	22	100.00
		其他	0	100.00

由图 10-12 可以看出：样本二（男生抽样全体）月生活费 400～500 元所占频数最高，是月生活费的众数。分析众数后，我们进一步分析月生活费的平均水平，得出结果（表 10-11）。

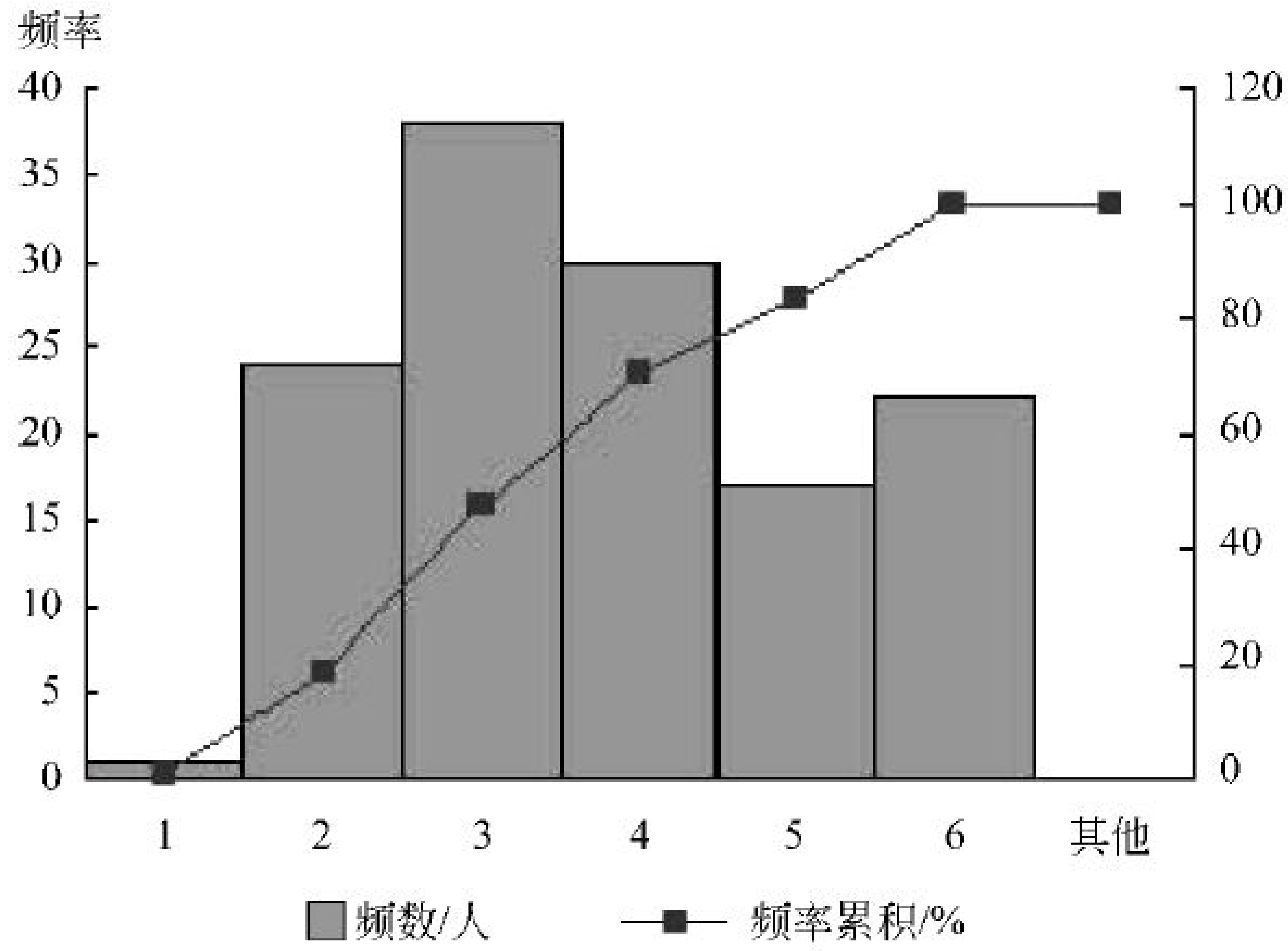


图 10-12　样本二月生活费分布直方图

表 10-11　样本二(男生)平均月生活费置信区间的构造表

月生活费水平/元	频数统计	组中值	描述统计输出结果	
300 以下	1	250	样本数据个数	132
300～400	24	350	样本标准差	135.93
400～500	38	450	样本均值	528.79
500～600	30	550	置信水平	95%
600～700	17	650	自由度	131
700 以上	22	750	t 值	1.978 237
			抽样平均误差	11.831 19
			误差范围	23.404 9
			置信下限	505.385 1
			置信上限	552.194 9

从上述分析可知：我们有 95%的把握认为人大本科男生的月生活费平均水平在 505.39～552.19 元。

(三) 对样本三的分析

由整理后输入计算机的数据，绘制出样本三月生活费水平的频数分布表(表 10-12)和直方图(图 10-13)，结果如下。

表 10-12　样本三月生活费水平的频数分布表

序号	月生活费水平/元	接收	频数/人	频率累积/%
1	300 以下	1	3	2.26
2	300～400	2	17	15.04
3	400～500	3	36	42.11
4	500～600	4	32	66.17
5	600～700	5	16	78.20
6	700 以上	6	29	100.00
		其他	0	100.00

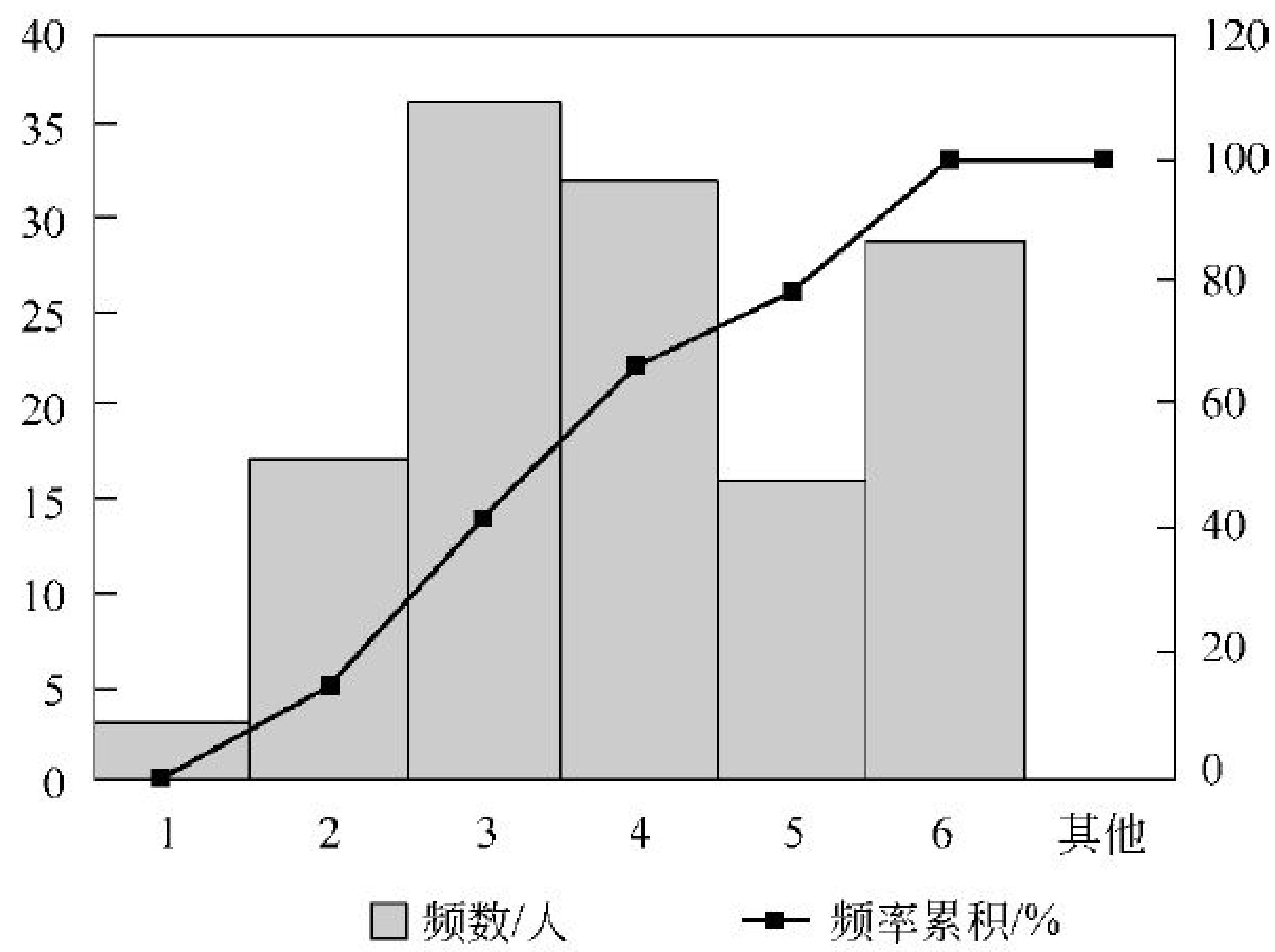


图 10-13 样本三月生活费分布直方图

由图 10-13 可以看出：样本三(女生抽样全体)月生活费 400～500 元所占频数最高。我们进一步分析月生活费的平均水平，得出结果(表 10-13)。

表 10-13 样本三(女生)平均月生活费置信区间的构造表

月生活费水平/元	频数统计/人	组中值/元	描述统计输出结果	
300 以下	3	250	样本数据个数	133
300～400	17	350	样本标准差	141.1
400～500	36	450	样本均值	546.24
500～600	32	550	置信水平	95%
600～700	16	650	自由度	132
700 以上	29	750	t 值	1.977 96
			抽样平均误差	12.234 92
			误差范围	24.200 19
			置信下限	522.039 8

从上述分析可知：我们有 95% 的把握认为人大本科生女生的月生活费平均水平在 522.04～570.44 元。

(四) 对各个年级的分析

1. 对大一平均月生活费水平的分析

由整理后输入计算机的数据，绘制出大一月生活费水平的频数分布表(表 10-14)和直方图(图 10-14)，结果如下。

由图 10-14 可以看出：大一抽样全体的月生活费众数为 400～500 元。

表 10-14 大一月生活费水平的频数分布表

序号	月生活费水平/元	接收	频数	频率累积/%
1	300	1	1	1.75
2	300～400	2	9	17.54
3	400～500	3	19	50.88
4	500～600	4	15	77.19
5	600～700	5	5	85.96
6	700 以上	6	8	100.00
		其他	0	100.00

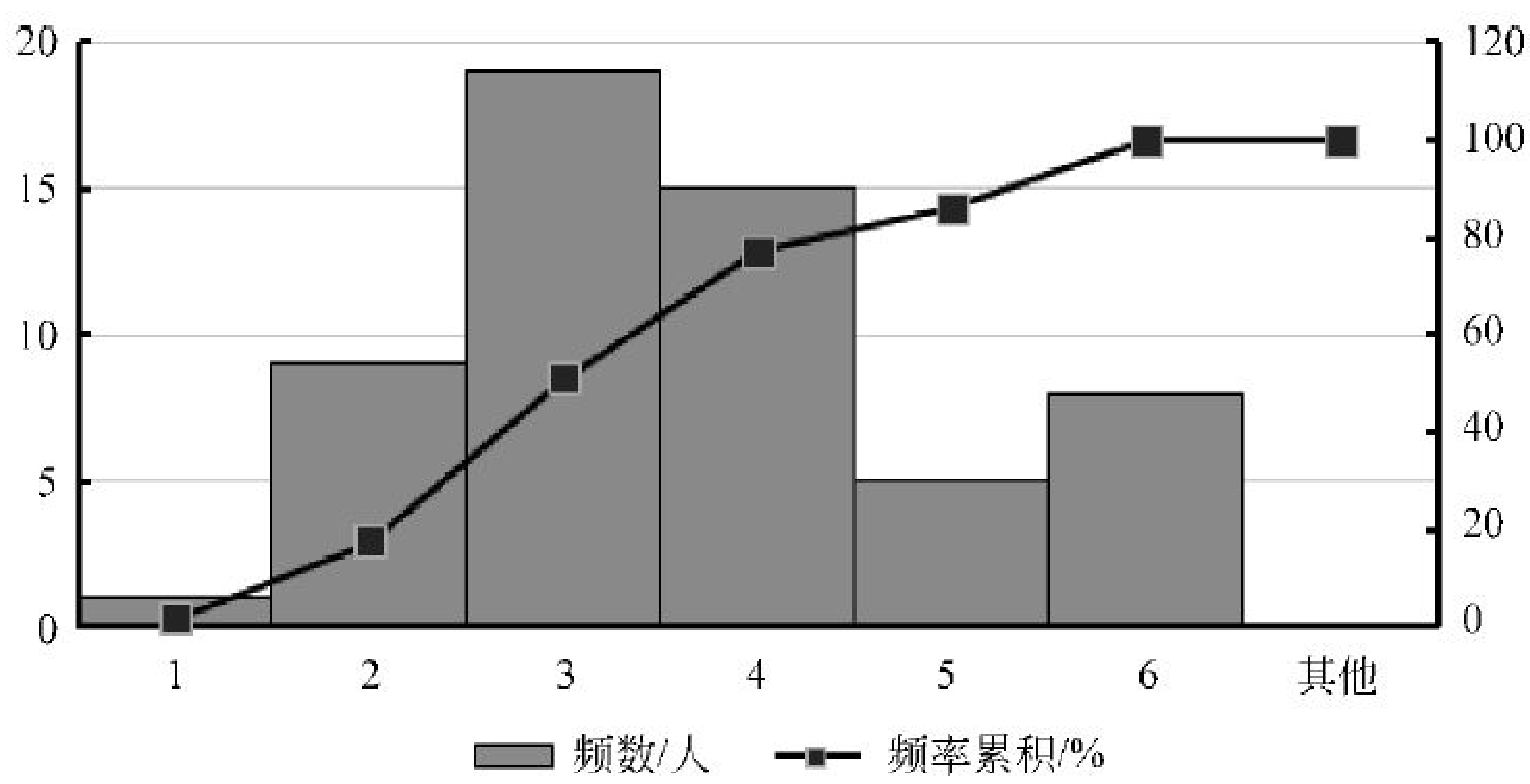


图 10-14 大学一年级月生活费分布直方图

我们进一步分析月生活费的平均水平，得出结果(表 10-15)：

表 10-15 大一平均月生活费置信区间的构造表

月生活费水平/元	组中值	频数	抽样平均误差	17.221 344 54
300 以下	250	1	置信水平	0.95
300～400	350	9	自由度	56
400～500	450	19	t 值	2.003 239 388
500～600	550	15	误差范围	34.498 475 68
600～700	650	5	置信下限	482.168 224 3
700 以上	750	8	置信上限	551.165 175 7
			样本均值	516.666 7
			样本标准差	130.018 3

从上述分析可知：我们有 95%的把握认为大一的月生活费平均水平在 522.04～570.44 元。

2. 对大二平均月生活费水平的分析

由整理后输入计算机的数据，绘制出大二月生活费水平的频数分布表(表 10-16)和

直方图(图 10-15),结果如下。

表 10-16　大二月生活费水平的频数分布表

序号	月生活费水平/元	接收	频数/人	频率累积/%
1	300 以下	1	1	1.35
2	300～400	2	10	14.86
3	400～500	3	16	36.49
4	500～600	4	20	63.51
5	600～700	5	13	81.08
6	700 以上	6	14	100.00
		其他	0	100.00

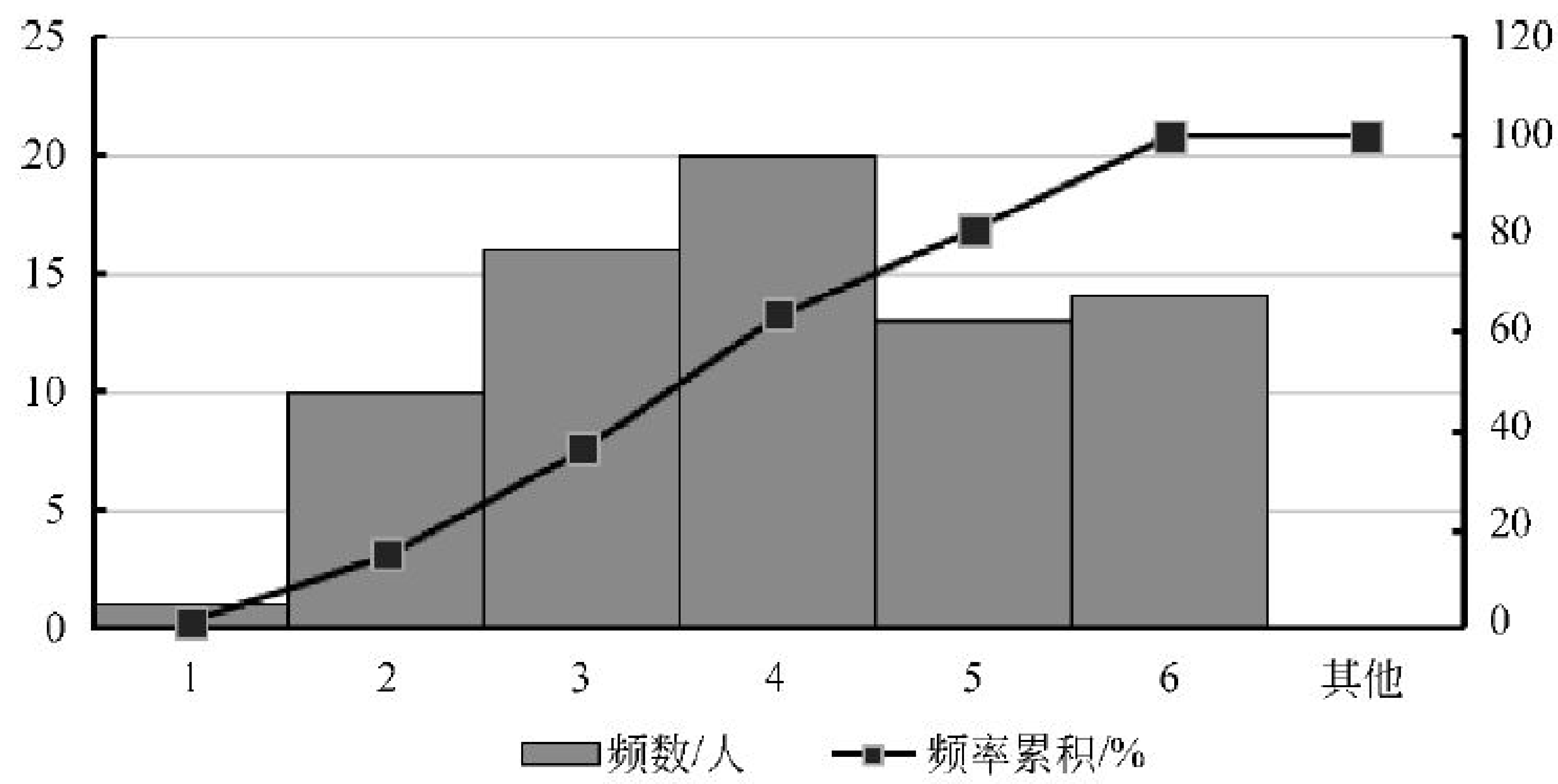


图 10-15　大学二年级月生活费分布直方图

由图 10-15 可以看出：大二抽样全体的月生活费众数为 500～600 元。

我们进一步分析月生活费的平均水平,得出结果(表 10-17)。

表 10-17　大二平均月生活费置信区间的构造表

月生活费水平/元	组中值/元	频数/人	抽样平均误差	15.746 626 15
300 以下	250	1	置信水平	0.95
300～400	350	10	自由度	73
400～500	450	16	t 值	1.992 998 477
500～600	550	20	误差范围	31.383 001 94
600～700	650	13	置信下限	521.319 698 1
700 以上	750	14	置信上限	584.085 701 9
		74	样本均值	552.702 7
			样本标准差	135.457 6

从上述分析可知：我们有 95% 的把握认为大二的月生活费平均水平在 521.32～584.09 元。

3. 对大三平均月生活费水平的分析

由整理后输入计算机的数据,绘制出大三月生活费水平的频数分布表(表 10-18)和直方图(图 10-16),结果如下。

表 10-18　大三月生活费水平的频数分布表

序号	月生活费水平/元	接收	频数/人	频率累积/%
1	300 以下	1	1	1.41
2	300～400	2	14	21.13
3	400～500	3	21	50.70
4	500～600	4	12	67.61
5	600～700	5	10	81.69
6	700 以上	6	13	100.00
		其他	0	100.00

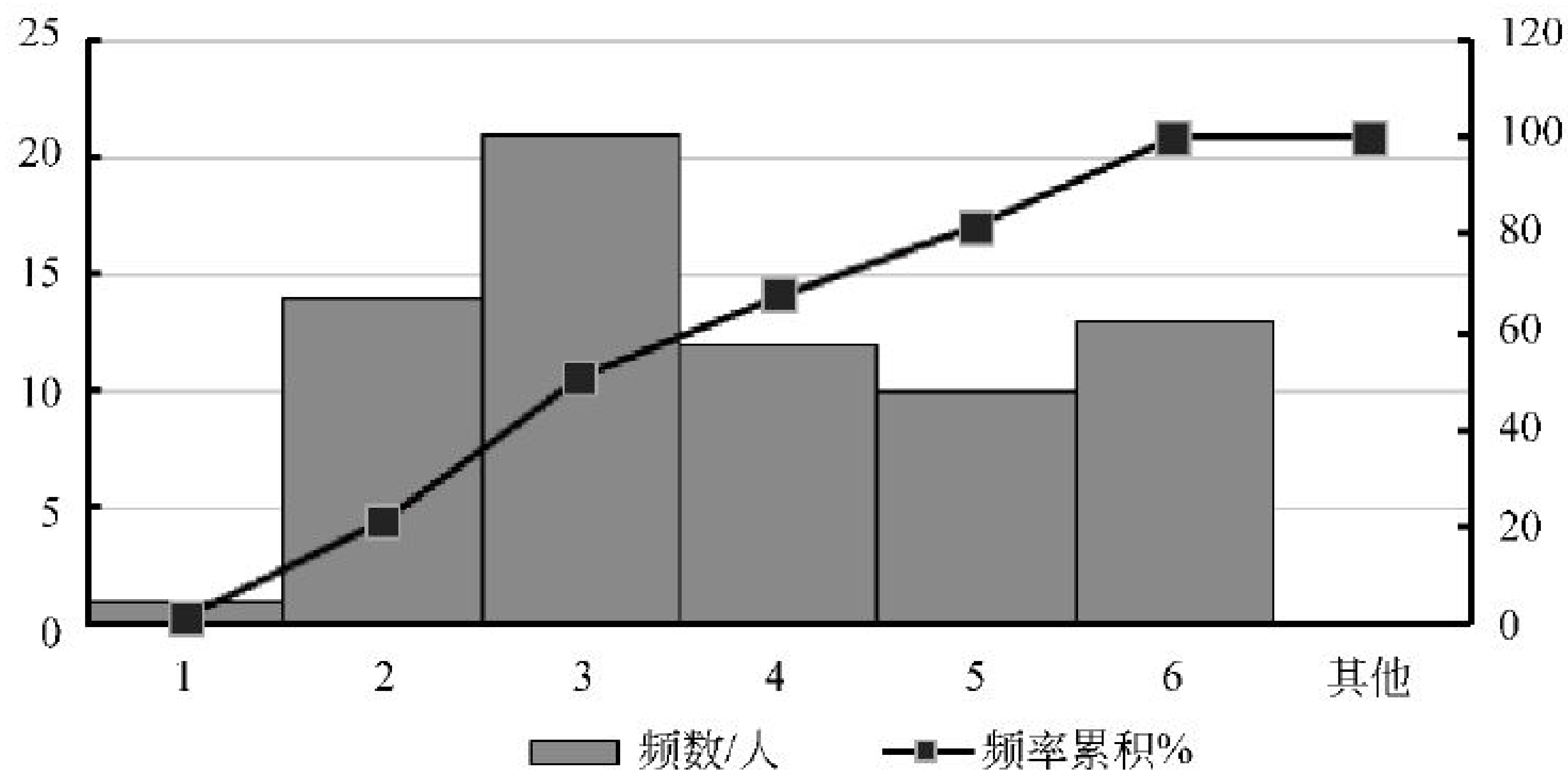


图 10-16　大学三年级月生活费分布直方图

由图 10-16 可以看出：大三抽样全体的月生活费众数为 400～500 元。

我们进一步分析月生活费的平均水平,得出结果(表 10-19)。

表 10-19　大三平均月生活费置信区间的构造表

月生活费水平/元	组中值/元	频　数	抽样平均误差	18.102 507 64
300 以下	250	1	置信水平	0.95
300～400	350	8	自由度	62
400～500	450	18	t 值	1.998 969 31
500～600	550	15	误差范围	36.186 357 22
600～700	650	5	置信下限	513.813 642 8
700 以上	750	16	置信上限	586.186 357 2
		63	样本均值	550
			样本标准差	143.684 2

从上述分析可知：我们有 95% 的把握认为大三的月生活费平均水平在 513.81～

586.19 元。

4. 对大四平均月生活费水平的分析

由整理后输入计算机的数据，绘制出大四月生活费水平的频数分布表（表 10-20）和直方图（图 10-17），结果如下。

表 10-20　大四月生活费水平的频数分布表

序号	月生活费水平/元	接收	频数/人	频率累积/%
1	300 以下	1	1	1.41
2	300～400	2	14	21.13
3	400～500	3	21	50.70
4	500～600	4	12	67.61
5	600～700	5	10	81.69
6	700 以上	6	13	100.00
		其他	0	100.00

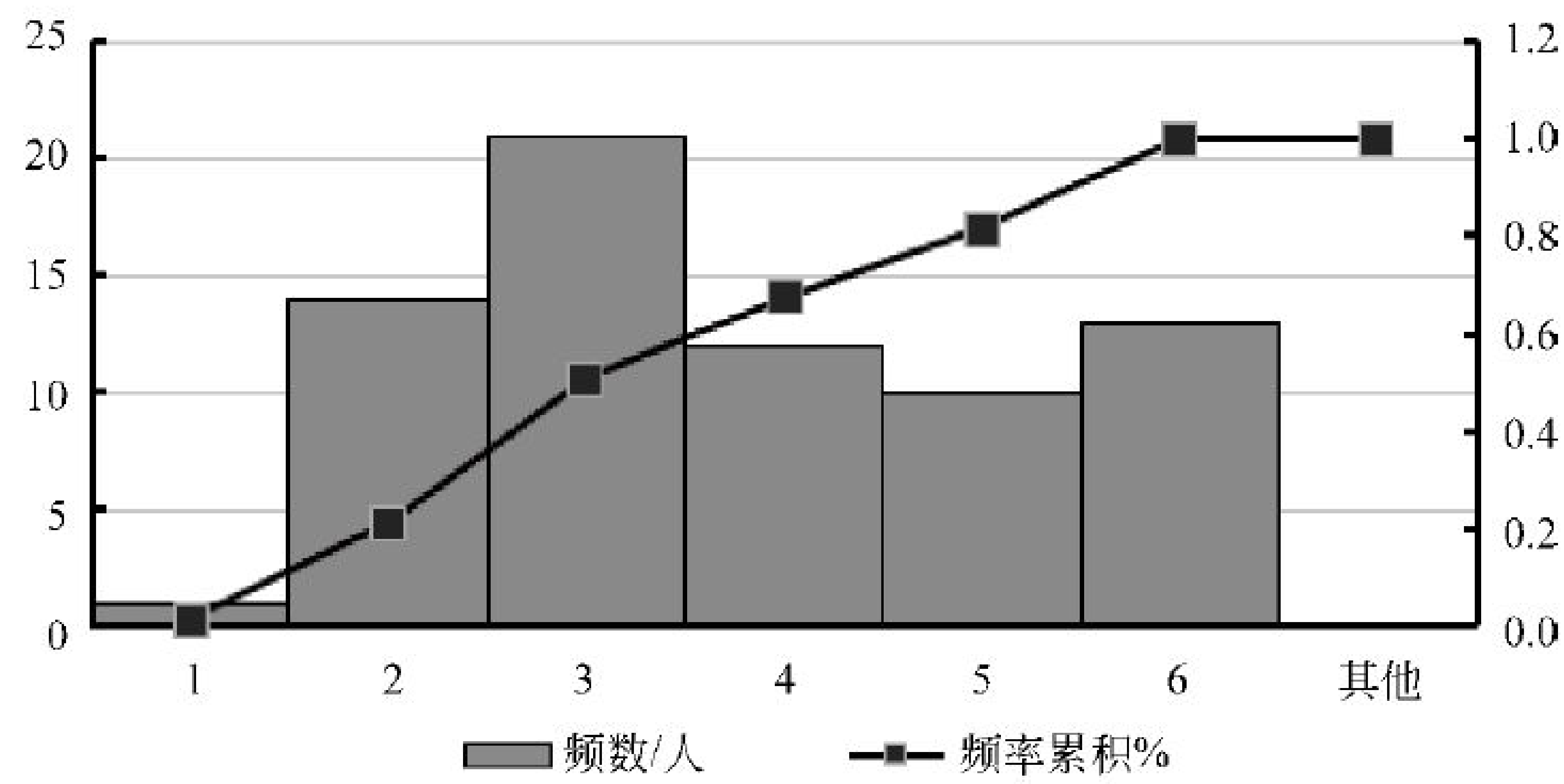


图 10-17　大学四年级月生活费分布直方图

由图 10-17 可以看出：大四抽样全体的月生活费众数为 400～500 元。

我们进一步分析月生活费的平均水平，得出结果（表 10-21）。

表 10-21　大四平均月生活费置信区间的构造表

月生活费水平/元	组中值/元	频数/人	抽样平均误差	17.044 997 28
300 以下	250	1	置信水平	0.95
300～400	350	14	自由度	70
400～500	450	21	t 值	1.994 435 479
500～600	550	12	误差范围	33.995 147 32
600～700	650	10	置信下限	493.469 652 7
700 以上	750	13	置信上限	561.459 947 3
		71	样本均值	527.464 8
			样本标准差	143.623 7

从上述分析可知：我们有 95％的把握认为大四的月生活费平均水平在 493.47～561.46 元。

5. 关于大学生生活费均值双因素方差分析

在对各个样本生活费均值进行区间估计和描述统计的基础上，我们进一步对影响生活费均值的各个因素(性别、年级)进行方差分析，见表 10-22。具体分析过程如下。

表 10-22 影响生活费均值的各个因素

性　别	大一	大二	大三	大四
男	496.153 8	555.882 4	556.451 61	506.097 56
女	533.871 0	550.000 0	543.750 00	556.66 667

建立假设：

对于因素 A：H_0：$\mu_1=\mu_2$　性别对生活费均值无影响；

H_1：$m_1\neq m_2$，性别对生活费均值有影响。

对于因素 B：H_0：$\mu_1=\mu_2=\mu_3=\mu_4$，年级对生活费均值无影响；

H_1：μ_1，μ_2，μ_3，μ_4，不全相等，年级对生活费均值有影响。

Excel 输出的结果如下：

方差分析：无重复双因素分析

SUMMARY	计数	求　和	平　均	方　差	
行 1	4	2 114.585	528.646 33	1 026.381 972	
行 2	4	2 184.288	546.071 91	93.977 677 81	
列 1	2	1 030.025	515.012 385	711.292 456 4	
列 2	2	1 105.882	552.941 175	17.301 020 76	
列 3	2	1 100.202	550.100 805	80.665 448 3	
列 4	2	1 062.764	531.382 115	1 278.617 443	

方差分析

差异源	SS	df	MS	*F*	*P*-value	*F* crit
行	607.301 7	1	607.301 676 7	1.230 539 087	0.348 229 816	10.127 96
列	1 880.504	3	626.834 752 4	1.270 117 791	0.424 428 733	9.276 619
误差	1 480.575	3	493.524 897 3			
总计	3 968.381	7				

由表可知，对于因素 A，因为：$FA<F\quad crit$，故接受原假设，说明性别对生活费没有显著影响；

对于因素 B，因为：$FB<F\quad crit$，故接受原假设，说明年级对生活费没有显著影响。

二、对生活费来源的分析

(一) 对抽样总体的分析

对抽样总体的分析如表 10-23 和图 10-18 所示。

表 10-23　样本一生活费主要来源

接　收	频　数	频率/%
1. 父母	229	86.42
2. 勤工俭学	21	7.92
3. 助学贷款	12	4.53
4. 其他	3	1.13
合计	265	100.00

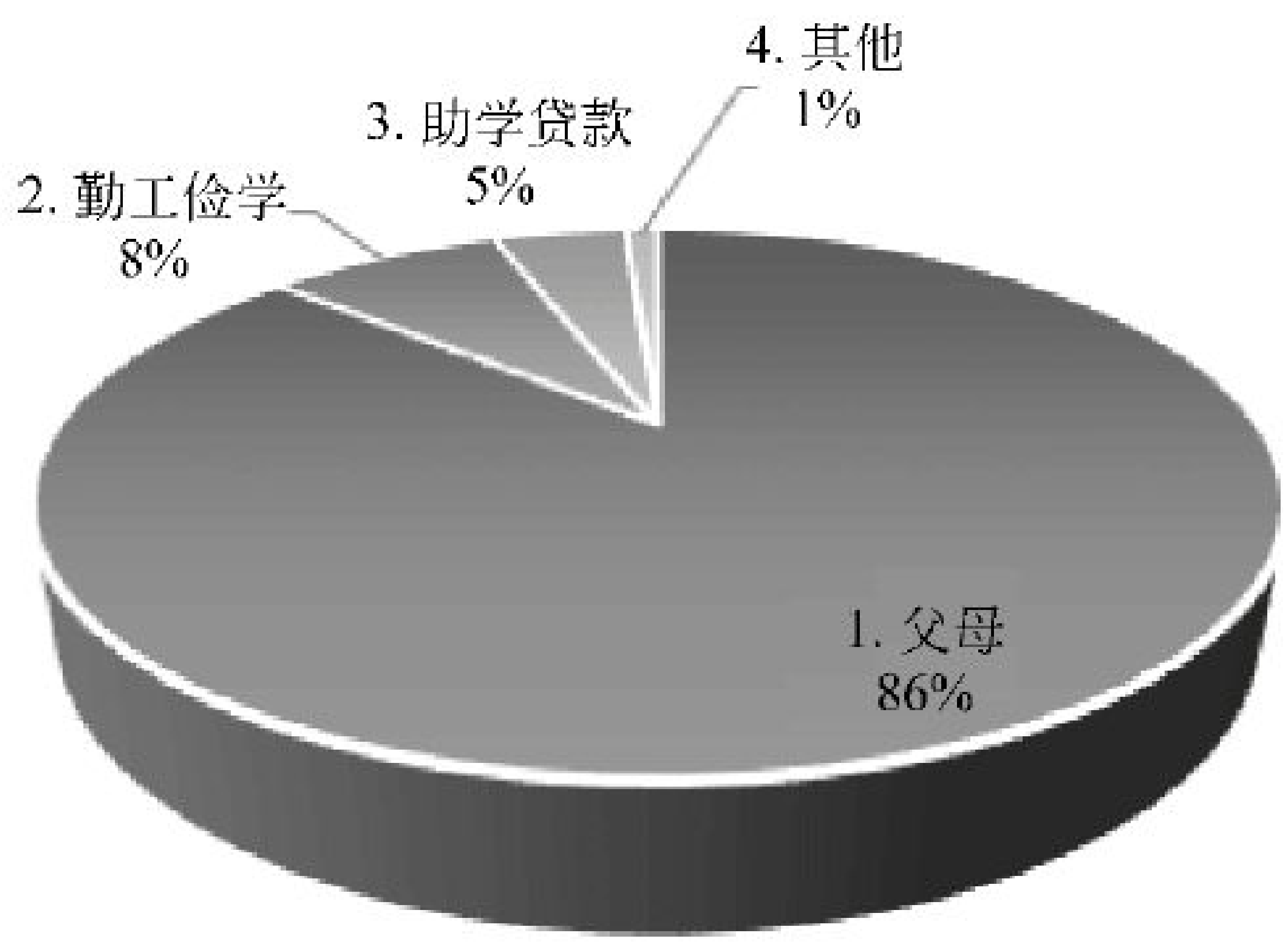


图 10-18　样本一生活费主要来源——饼图

由图 10-18 可以看出：样本一的生活费主要来源中，父母所占频数比例最高，为 86%。

（二）对样本二、样本三的分析

样本二、样本三的分析如表 10-24、表 10-25 和图 10-19 所示。

表 10-24　样本二生活费主要来源——频数分布

接　收	频　数	频率%
1. 父母	110	83
2. 勤工俭学	14	11
3. 助学贷款	5	4
4. 其他	3	2
合计	132	100.00

表 10-25　样本三生活费主要来源——频数分布

接　收	频　数	频率/%
1. 父母	120	90
2. 勤工俭学	6	5
3. 助学贷款	7	5
4. 其他	0	0
合计	133	100.00

由图 10-19 可以看出：样本二(男生)生活费的主要来源中，父母所占频数比例最高，为 83%；与此相比，样本三(女生)生活费主要来源中所占频数比例最高的父母一项的比

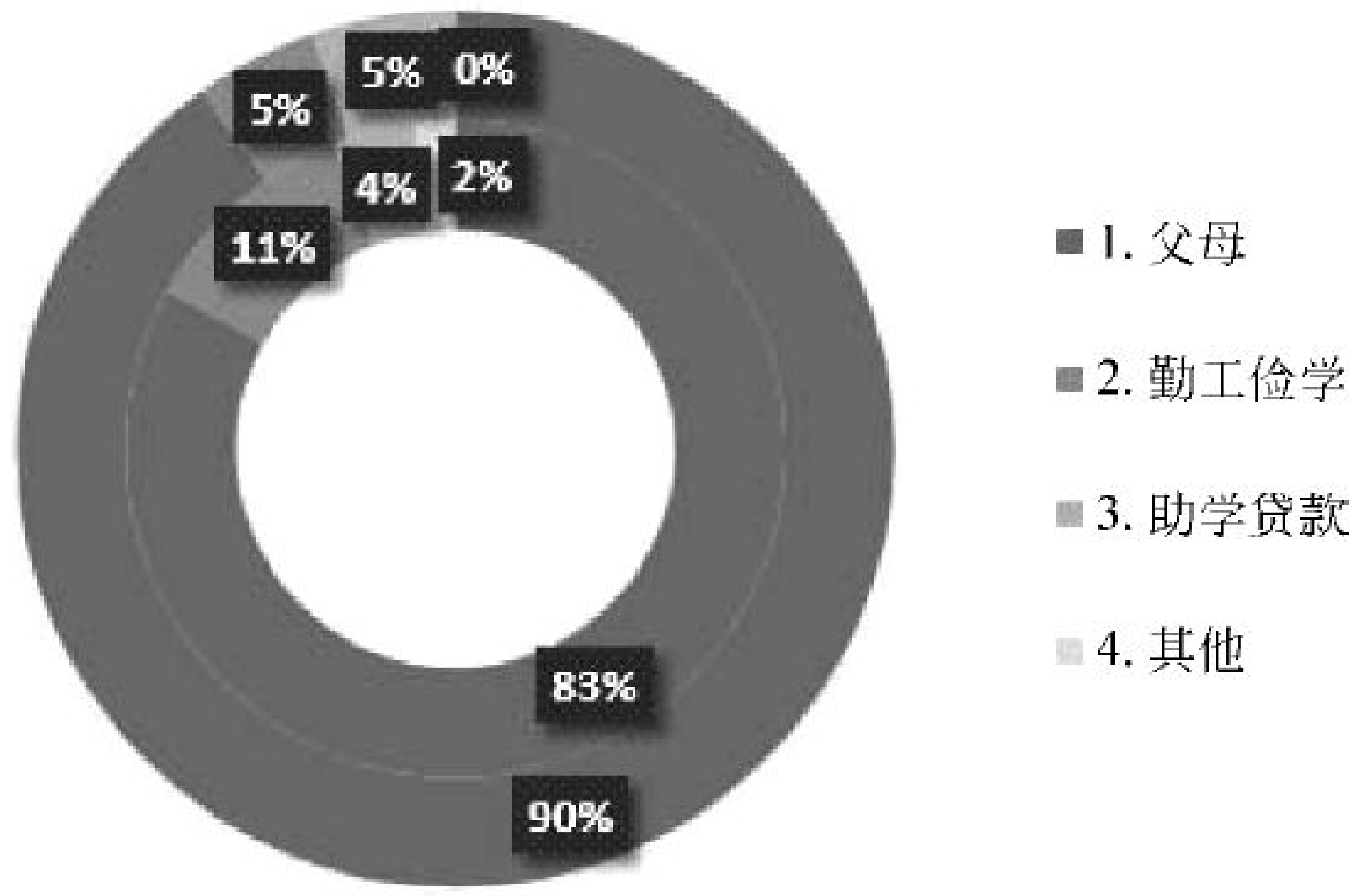


图 10-19　样本二与样本三生活费主要来源——对照图

注：内环为样本二，外环为样本三

例为 90%，此外，样本二(男生)勤工俭学频数比例(11%)较样本三(女生)频数比例(5%)高出许多；样本二(男生)其他源频数比例(2%)较样本三(女生)其他源频数比例(0%)高。

(三) 年级收入来源的分析

年级收入来源的分析如表 10-26～表 10-29、图 10-20 所示。

表 10-26　大一收入来源

来　源	频　数	频率/%
1. 父母	55	96
2. 勤工俭学	1	2
3. 助学贷款	1	2
4. 其他	0	0
合　计	57	100%

表 10-27　大二收入来源

来　源	频　数	频率/%
1. 父母	65	88
2. 勤工俭学	5	7
3. 助学贷款	4	5
4. 其他	0	0
合　计	74	100%

表 10-28　大三收入来源

来　源	频　数	频率/%
1. 父母	56	89
2. 勤工俭学	2	3
3. 助学贷款	4	6
4. 其他	1	2
合　计	63	100

表 10-29　大四收入来源

来　源	频　数	频率/%
1. 父母	53	75
2. 勤工俭学	13	18
3. 助学贷款	3	4
4. 其他	2	3
合　计	71	100

由图 10-20 可以看出：各个年级的同学生活费的大部分来自父母，其中大一年级同学生活费来自父母的比例最高，占到了 96.49%；大四年级同学生活费来自父母的比例最低，为 74.65%。另外，勤工俭学也成为学生生活费来源中不可忽视的一项，在大四年级中所占比例最高，达到了 18.31%，也就是 1/5；其次是大二年级，所占比例为 6.75%，比例最低的为大一年级，只占到了生活费比例总数的 1.96%。伴随着助学贷款政策的逐步深入，在大学生生活费中所占的比例也令人瞩目，在各年级学生生活费中所占比例由大到小依次为：大三 6.35%、大二 5.41%、大四 4.22%、大一 1.75%。值得注意的是，学生生

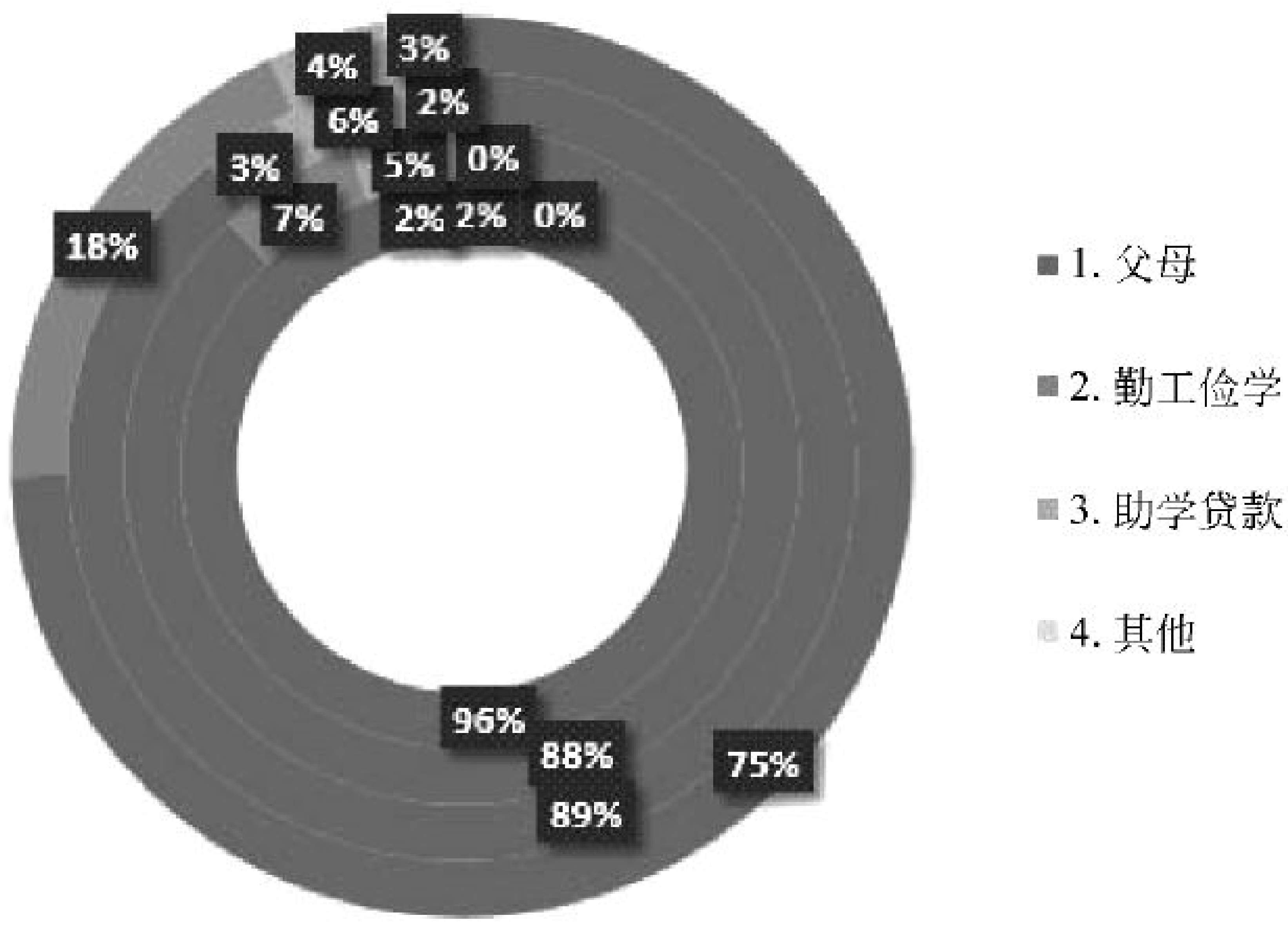


图 10-20　四个年级生活费主要来源——对照图

（注：从内到外依次是大一、大二、大三、大四）

活费来源，除了我们所估计的父母、勤工俭学、助学贷款三项外，还有其他来源，并且随年级的升高，所占比例也越来越大，分别为大四 2.82%、大三 1.59%、大二 0、大一 0。

三、对生活费主要支出结构的分析

（一）对不同性别的生活费主要支出的分析

1. 对样本一的分析（表 10-30、图 10-21）

表 10-30　样本一的生活费主要支出结构

项　目	接收	频率	累积/%
伙食费	1	228	86.04
衣着	2	18	92.83
学习用品	3	8	95.85
日化用品	4	1	96.23
娱乐休闲	5	9	99.62
其他	6	1	100.00
合计		0	100.00

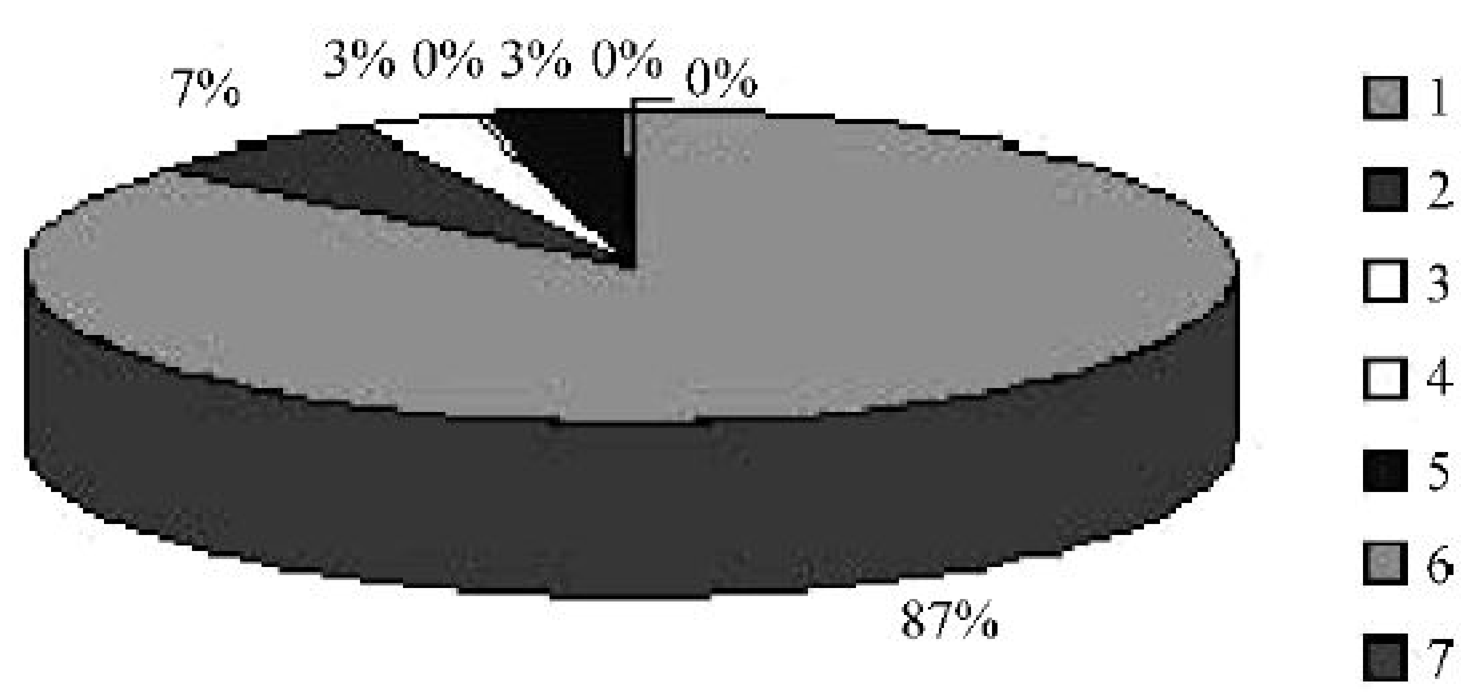


图 10-21　生活费支出结构图

由表 10-30 和图 10-21 可以看出：样本一生活费主要支出中，伙食费所占频数比例最高，为 87%。

2. 对样本二和样本三的分析(表 10-31、表 10-32、图 10-22)。

表 10-31 样本二生活费主要支出——频率分布

接 收	频数	频率/ %
1. 伙食费	120	91
2. 衣着	5	4
3. 学习用品	4	3
4. 日化用品	0	0
5. 娱乐休闲	3	2
6. 其他	0	0
合 计	132	100.00

表 10-32 样本三生活费主要支出——频率分布

接 收	频数	频率/ %
1. 伙食费	108	81
2. 衣着	13	10
3. 学习用品	4	3
4. 日化用品	1	1
5. 娱乐休闲	6	5
6. 其他	1	1
合 计	0	100.00

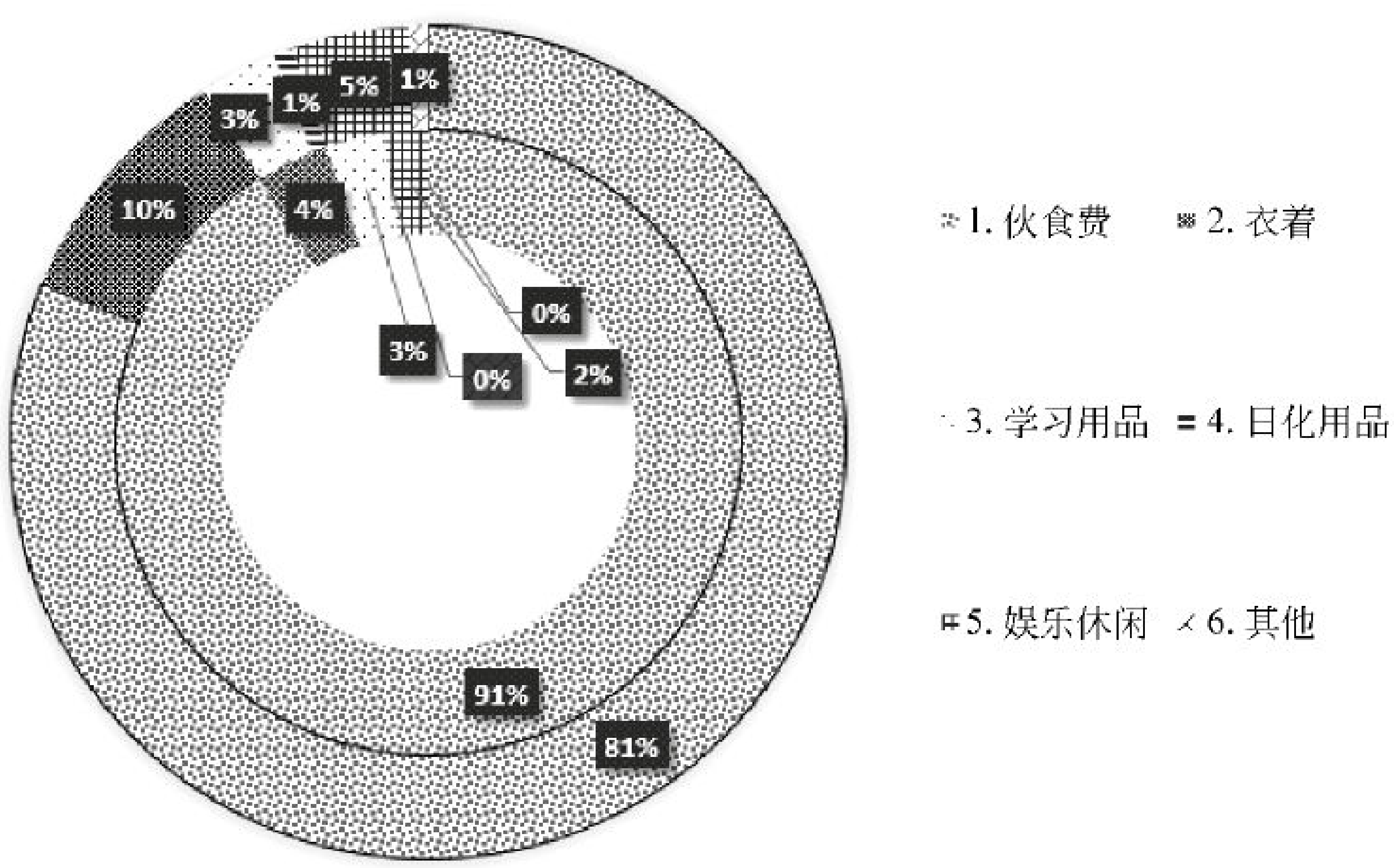


图 10-22 样本二与样本三生活费主要开支——对照环形图

(注：外环为样本三,内环为样本二。)

由图 10-22 可以看出：样本二(男生)生活费的主要支出中,伙食费所占频数比例最高为 91%；同时,样本三(女生)生活费中伙食费所占频数比例最高为 80%,此外,样本三(女生)的衣着费用频数比例 4%。

(二) 对不同年级生活费主要支出的分析(表 10-33～表 10-36、图 10-23)

表 10-33 大一生活费主要支出——频数分布表

支出项目	频数	频率累积/%
伙食费	52	91.23
衣着	3	96.49
学习用品	0	96.49
日化用品	0	96.49
娱乐休闲	2	100.00
其他	0	100.00
合 他	57	100.00

表 10-34 大二生活费主要支出——频数分布表

支出项目	频数	频率累积/%
伙食费	66	89.19
衣着	3	93.24
学习用品	3	97.30
日化用品	0	97.30
娱乐休闲	2	100.00
其他	0	100.00
合 计	74	100.00

表 10-35　大三生活费主要支出——频数分布表

支出项目	频数	频率累积 /%
伙食费	59	93.65
衣着	1	95.24
学习用品	2	98.41
日化用品	1	100.00
娱乐休闲	0	100.00
其他	0	100.00
合　计	63	100.00

表 10-36　大四生活费主要支出——频数分布表

支出项目	频数	频率累积 /%
伙食费	51	71.83
衣着	11	87.32
学习用品	3	91.55
日化用品	0	91.55
娱乐休闲	5	98.59
其他	1	100.00
合　计	71	100.00

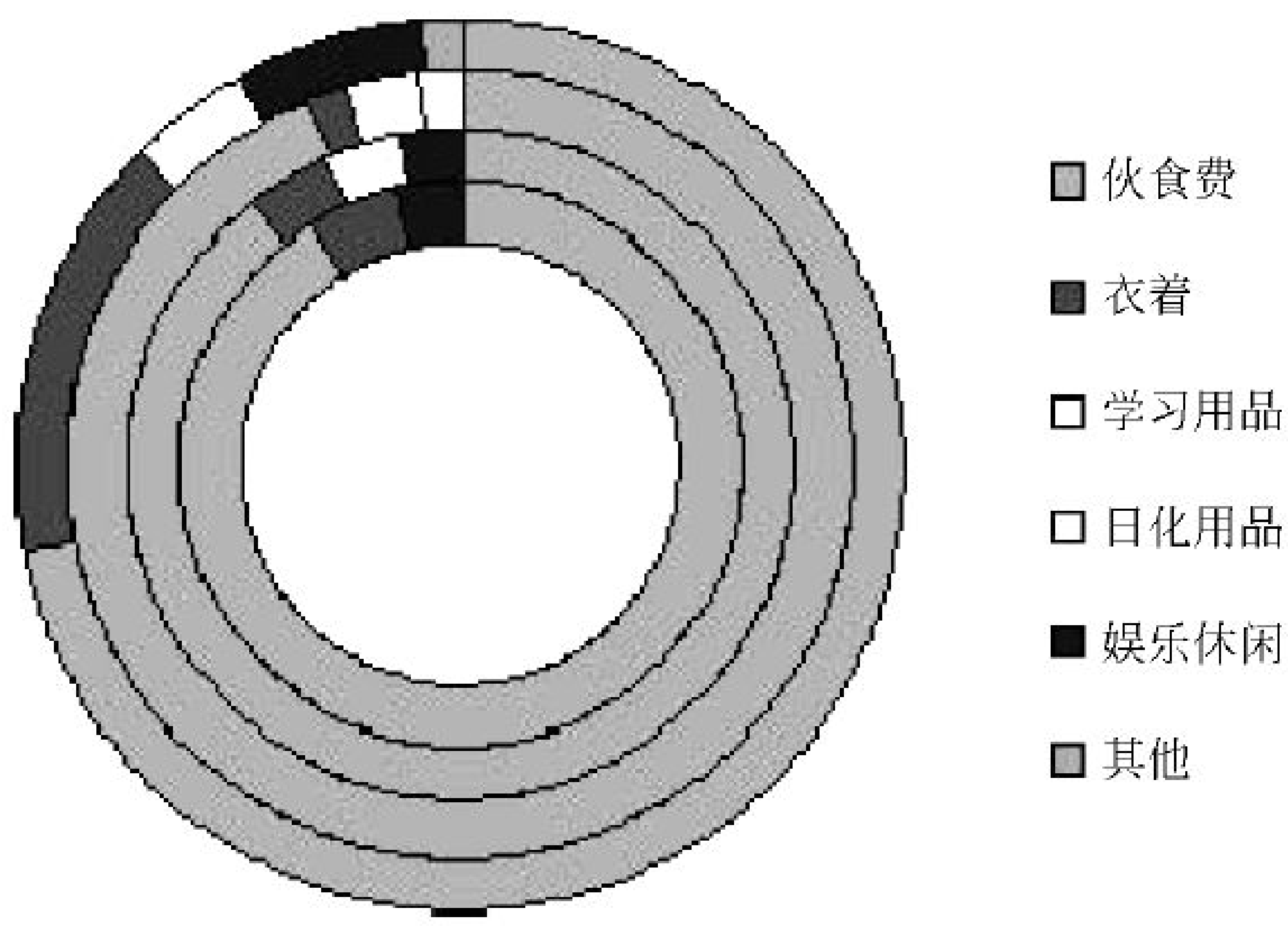


图 10-23　各年级主要开支结构

（注：从内到外依次为大一、大二、大三、大四。）

从图 10-23 可以看出，在调查对象中，四个年级以伙食费为主要开支的人数分别占各个年级样本容量的 91.23%、89.19%、93.65%、71.83%。大四学生中以衣着为主要开支的人数比例在四个年级中占到最多，达到 15.49%

四、总结

通过以上的统计分析，我们基本得出以下结论。

（一）对于生活费均值的分析结果

（1）通过对人大全体本科生生活费均值的区间估计，发现全体本科生总体的生活费均值在 520.79～554.31 元。

（2）各个分类总体而言，我们以不同年级、不同性别为分类标准对总体又进行了划分，估计出全体本科生中各个年级以及男女生分别的生活费均值。方差分析的结果表明，年级以及性别对生活费均值没有显著影响，即不同年级、不同性别的生活费基本一致。

（二）对于生活费来源的分析结果

（1）描述统计的结果显示，生活费的主要来源都集中在父母供给中，其他来源依次是：勤工俭学、助学贷款及其他。

（2）男女生而言，男生的生活费主要来源中来自父母的比女生稍低，而勤工俭学的比

女生稍多,助学贷款比例相差不大。

(3) 不同年级而言,高年级生活费来自父母的比例比低年级的稍低,而勤工俭学的比例比低年级的稍高,助学贷款比例相差不大。

(4) 以上结果表明助学贷款政策的实施面不够广,人大本科生还是传统的以靠父母读书为主。

(三) 对于生活费主要支出的分析结果

(1) 就抽样总体而言,生活费的主要支出集中在伙食费上,其他支出依次是:衣着、娱乐休闲、学习用品、日化用品。

(2) 男女生而言,男生中以伙食费为主要支出的比例比女生高,而女生中以衣着为主要支出的比例比男生高。

(3) 不同年级而言,大四学生中以衣着为主要支出的比例明显高过其他年级。

(4) 由恩格尔系数表明,学生伙食费占支出的绝大部分,学生的生活刚刚达到温饱水平而已。

五、附录

每一行表示一张问卷结果,见表 10-37。

表 10-37 调查的原始数据

序号	性别	年级	生活费	主要来源 1	主要来源 2	主要来源 3	主要来源 4	主要开支 1	主要开支 2	主要开支 3
1	1	1	3	1	2	3	4	1	2	3
2	1	1	4	1	2			1	2	5
3	1	1	6	1				1	2	5
4	1	1	4	1				1	5	3
5	1	1	3	1				1	3	2
6	1	1	3	1				1	2	3
7	1	1	6	1	4			5	1	
8	1	1	3	1	4			1	3	2
9	1	1	4	1	2	3		2	1	3
10	1	1	4	1	2	4		1	5	6
11	1	1	3	1				2	5	1
12	1	1	4	1				1	3	5
13	1	1	3	1	2			1	3	2
14	1	1	2	1	2			1	3	4
15	1	1	2	1				1	3	4
16	1	1	4	1				1	3	2
17	1	1	3	1	4			1	3	2
18	1	1	2	1	4			1	3	5
19	1	1	5	1	4			1	3	5
20	1	1	2	1	4			1	3	4
21	1	1	3	1	2			1	5	2
22	1	1	4	1	2	4	3	1	3	4
23	1	1	3	1				1	2	3

续表

序号	性别	年级	生活费	主要来源 1	主要来源 2	主要来源 3	主要来源 4	主要开支 1	主要开支 2	主要开支 3
24	1	1	4	1	2			1	2	3
25	1	1	4	1				1	4	2
26	1	1	2	1	3			1	2	3
27	1	2	5	1	3	2	4	1	2	
28	1	2	3	1	4	3	2	1		
29	1	2	4	2	3	1	4	1	5	3
30	1	2	2	1	3	2		1	2	3
31	1	2	3	1	2			1	2	3
32	1	2	2	1	3			1	3	4
33	1	2	2	1	4	2	3	1	2	3
34	1	2	6	1	4			1	5	
35	1	2	4	3	1	2		1	6	
36	1	2	5	2	1	4		1	2	
37	1	2	2	1	2			1	2	3
38	1	2	5	1	2	4	3	1	2	3
39	1	2	6	1	2			1	5	2
40	1	2	3	1	2	3	4	1		
41	1	2	6	1				1	5	6
42	1	2	3	2	1	3	4	3	1	
43	1	2	5	1				1	2	6
44	1	2	3	1	2	4	3	1	2	3
45	1	2	3	1	2	3		1		
46	1	2	6	1				1	5	
47	1	2	5	1	3	2		1		
48	1	2	3	1				1		
49	1	2	4	1				1	6	3
50	1	2	4	1				5	1	2
51	1	2	5	1	3			1	2	5
52	1	2	6	1				1	2	6
53	1	2	4	1				1	2	3
54	1	2	4	1	2	4	3	1	2	4
55	1	2	5	1	3	2	4	1	2	5
56	1	2	4	1	3			1	2	3
57	1	2	5	1				1	2	5
58	1	2	5	1	4			1	5	
59	1	2	2	2	1	3	4	1	2	3
60	1	2	4	1	4			1	5	3
61	1	3	6	1	2	4		1	3	2
62	1	3	5	2	1			1	5	6
63	1	3	3	1	2			1	3	2

续表

序号	性别	年级	生活费	主要来源 1	主要来源 2	主要来源 3	主要来源 4	主要开支 1	主要开支 2	主要开支 3
64	1	3	6	2				1	3	5
65	1	3	3	1				1		
66	1	3	4	1	3	2	4	1	5	6
67	1	3	2	4	2	3	1	1		
68	1	3	4	1	2	4	3	1		
69	1	3	4	1	2			1		
70	1	3	4	3	1	2		1	5	
71	1	3	6	1	2			1	2	3
72	1	3	6	1	2	4	3	1	5	2
73	1	3	6	1	2			1	2	5
74	1	3	6	1	2			1	5	6
75	1	3	5	1	2	4		1	5	3
76	1	3	5	1				1		
77	1	3	1	1	2	3		1	3	4
78	1	3	3	1	2			1	3	2
79	1	3	3	1	2	4	3	1		
80	1	3	4	1	3	2		1	2	3
81	1	3	4	1	4			1		
82	1	3	2	1				1	5	2
83	1	3	4	1	2	3		1	2	
84	1	3	6	1	2			1	2	3
85	1	3	2	3	2			1	2	3
86	1	3	3	1	2	3	4	1	2	3
87	1	3	3	1	3	2	4	1	3	2
88	1	3	3	1	2	3	4	1	5	3
89	1	3	3	1				1	5	
90	1	3	4	3	1	2	4	1		
91	1	3	6	1				1	3	5
92	1	4	6	2	1	3	4	3	1	2
93	1	4	4	1	2	3	4	2	1	5
94	1	4	6	1	2			1	2	3
95	1	4	2	2	3	1	4	1	3	2
96	1	4	3	1	2	3	4	1	3	2
97	1	4	3	1	2			1	2	6
98	1	4	2	2	3	1		1	3	2
99	1	4	3	1				1	2	3
100	1	4	2	1	2	4	3	1	3	2
101	1	4	2	4				1	3	4
102	1	4	3	2	1	3		5	1	3
103	1	4	2	1	2	4	3	1	4	5

续表

序号	性别	年级	生活费	主要来源1	主要来源2	主要来源3	主要来源4	主要开支1	主要开支2	主要开支3
104	1	4	2	2	3	1	4	1	4	6
105	1	4	3	2	3	4	1	1	3	2
106	1	4	2	2	4	1		1	5	3
107	1	4	2	1	3	2	4	1	3	6
108	1	4	4	1				1	3	5
109	1	4	5	1	4			1	3	4
110	1	4	3	1				2	1	5
111	1	4	4	1	4	2	3	1	5	3
112	1	4	5	1	2	3		1	3	5
113	1	4	3	1	2	4	3	1	2	4
114	1	4	3	2	3	1		1	2	5
115	1	4	6	1				1	5	2
116	1	4	3	1	3			1	3	4
117	1	4	2	1	2	3		3	4	5
118	1	4	6	1				1	5	3
119	1	4	3	1	2			3	1	5
120	1	4	3	2	3	1		1	2	4
121	1	4	2	1	3	2	4	2	3	5
122	1	4	4	1	4	2	3	1	3	2
123	1	4	6	1	2			1	5	2
124	1	4	3	1				1	2	5
125	1	4	3	1	3	2	4	1	3	4
126	1	4	5	1	2			1	5	2
127	1	4	6	1				1	2	5
128	1	4	5	3	2	1		1	3	4
129	1	4	2	1	2	4	3	1	3	2
130	1	4	3	1	2			1	3	2
131	1	4	6	1				1	5	
132	1	4	4	4	2			1	2	3
133	2	1	1	1	2			1	3	
134	2	1	4	1	2	3		1	2	5
135	2	1	3	1	2	4		1	2	
136	2	1	3	1	2			1	3	
137	2	1	5	1				1	2	4
138	2	1	5	1				1	5	3
139	2	1	2	1	2	4	3	1	3	
140	2	1	2	1	2	4	3	1	3	
141	2	1	6	1	2	3		1	2	
142	2	1	6	1	2			1	2	6
143	2	1	3	1	2			1	2	3

续表

序号	性别	年级	生活费	主要来源 1	主要来源 2	主要来源 3	主要来源 4	主要开支 1	主要开支 2	主要开支 3
144	2	1	4	1	2			1		
145	2	1	4	1				5	1	3
146	2	1	5	1				1	6	2
147	2	1	3	1				1	2	
148	2	1	4	1	2	4		1	2	6
149	2	1	6	1	2			1	2	4
150	2	1	3	1				1		
151	2	1	3	1	2			1	2	
152	2	1	3	1				1	2	3
153	2	1	5	1	4	2		1	2	4
154	2	1	3	3	2	1		1	2	3
155	2	1	2	2	1	3		2	1	3
156	2	1	3	1	2			1	3	4
157	2	1	3	1	2			1	2	5
158	2	1	4	1				1	3	6
159	2	1	2	1				1	3	5
160	2	1	6	1	2			1	6	2
161	2	1	6	1				1		
162	2	1	4	1				1	6	
163	2	1	6	1				1	6	2
164	2	2	3	1				1	2	3
165	2	2	2	1	2	3		1	2	3
166	2	2	5	1				1	2	4
167	2	2	6	1				1	2	
168	2	2	3	3	1	4	2	1	3	4
169	2	2	4	3	1	2		1	2	4
170	2	2	4	1	3	2		1	2	5
171	2	2	3	3	1	2		1	2	3
172	2	2	5	1	2	3	4	1	2	4
173	2	2	6	1	2	3		3	2	4
174	2	2	6	1				2	1	5
175	2	2	6	1				2	4	1
176	2	2	6	1				1	2	5
177	2	2	6	1				1	2	5
178	2	2	2	1	2	4	3	1	2	5
179	2	2	4	1	2			1	2	3
180	2	2	4	1	2			2	1	4
181	2	2	6	1				1	2	4
182	2	2	3	1	2			1	3	4
183	2	2	6	1	3	4	2	1	2	4

续表

序号	性别	年级	生活费	主要来源1	主要来源2	主要来源3	主要来源4	主要开支1	主要开支2	主要开支3
184	2	2	4	1				1	4	5
185	2	2	5	1				1	2	3
186	2	2	3	2	1	3	4	1	3	4
187	2	2	4	1	2			1	3	4
188	2	2	3	1				1	2	4
189	2	2	4	1	2			5	1	3
190	2	2	3	1	2			1	2	3
191	2	2	3	1	3	2		1	3	2
192	2	2	3	1	2			1	2	4
193	2	2	2	1				1	2	3
194	2	2	1	1	2	3		1	3	2
195	2	2	4	1	4			1	4	3
196	2	2	5	1	2			1	2	3
197	2	2	4	1	2			1	5	2
198	2	2	4	1				1	2	4
199	2	2	4	1	2	3	4	1	5	3
200	2	2	2	1				1	3	6
201	2	2	2	1	2			1	3	2
202	2	2	4	1				1	3	4
203	2	2	6	1	2			3	2	1
204	2	3	6	1				1	2	
205	2	3	3	1	2	4	3	1	3	
206	2	3	3	1				1		
207	2	3	3	1	2			1	4	5
208	2	3	3	1	2			1	6	
209	2	3	4	1	2	4	3	1	3	5
210	2	3	3	1	2	4	3	1		
211	2	3	2	1	2			1	3	2
212	2	3	4	1	2			1	2	
213	2	3	3	1	2			1		
214	2	3	2	1	2			1		
215	2	3	6	1	2	3		1	2	
216	2	3	4	1				1	2	4
217	2	3	4	1	2	3		1		
218	2	3	3	1	2	4		1	2	3
219	2	3	6	1	4			1	3	
220	2	3	2	1	2			1	2	3
221	2	3	3	1	2	3	4	1	3	2
222	2	3	6	1				1	4	2
223	2	3	5	1				1	3	5
224	2	3	4	1	2	3		4	2	3
225	2	3	6	1				1	2	4

续表

序号	性别	年级	生活费	主要来源 1	主要来源 2	主要来源 3	主要来源 4	主要开支 1	主要开支 2	主要开支 3
226	2	3	4	1	2			1	2	4
227	2	3	2	1				3	4	2
228	2	3	5	1				1	4	5
229	2	3	2	1				1	2	4
230	2	3	6	1				3	4	1
231	2	3	6	1	2			2	3	1
232	2	3	3	1	2	4	3	1	2	3
233	2	3	4	1	2			1	4	3
234	2	3	3	3	1	2	4	1	3	4
235	2	3	6	1	2			1	4	5
236	2	4	6	1	2	4	3	1	2	4
237	2	4	3	2	1			1		
238	2	4	3	1	3	2	4	1	2	5
239	2	4	6	1				6	1	5
240	2	4	4	1				1	2	3
241	2	4	4	1	2	4		1		
242	2	4	5	1				1	2	6
243	2	4	6	2				5	2	4
244	2	4	6	1				1	2	4
245	2	4	4	1				1	2	4
246	2	4	4	2	4	1		5	2	1
247	2	4	3	2	3	1		1	3	2
248	2	4	5	1	2			5	1	4
249	2	4	3	1				2	3	1
250	2	4	4	1				1	3	6
251	2	4	2	1	2	4	3	2	1	3
252	2	4	3	3	1	2		1	4	3
253	2	4	5	1	2			1	2	4
254	2	4	1	3	2	1		1	3	2
255	2	4	4	1	2			2	1	3
256	2	4	3	1	2			2	1	4
257	2	4	3	1	2	3	4	1	4	3
258	2	4	5	1	2			1	2	5
259	2	4	6	1	2			2	5	1
260	2	4	2	1	2			2	4	1
261	2	4	6	1				5	3	1
262	2	4	5	1	2			1	5	4
263	2	4	4	1	2			2	4	3
264	2	4	5	1	2			1	4	6
265	2	4	2	1				2	1	3

(资料来源：贾俊平.《统计学》教学案例和教学项目汇编[M]. 北京：中国人民大学出版社，2004.)

思 考 题

1. 统计分析报告的类型有哪几种？
2. 统计分析报告的选题要求有哪些？
3. 统计分析报告的写作要求有哪几点？

练 习 题

1. 登录国家统计信息网，查阅资料，写一篇统计分析报告。
2. 了解一个企业，就企业数据写一篇统计分析报告。
3. 利用假期进行社会调查，写一篇调查分析报告。

参 考 文 献

[1] 凌明雁,柳秀春. 统计学[M]. 3 版. 北京：高等教育出版社,2012.
[2] 管于华. 统计学[M]. 3 版. 北京：高等教育出版社,2013.
[3] 徐国祥. 统计学[M]. 2 版. 上海：上海人民出版社,2014.
[4] 曾五一. 统计学导论[M]. 2 版. 北京：科学出版社,2016.
[5] 黄良文,朱建平. 统计学[M]. 3 版. 北京：中国统计出版社,2012.
[6] 贾俊平. 统计学[M]. 2 版. 北京：清华大学出版社,2006.
[7] 吴喜之. 统计学——从数据到结论[M]. 4 版. 北京：中国统计出版社,2013.
[8] 耿修林. 商务经济统计学[M]. 北京：科学出版社,2003.
[9] 曾五一. 统计学[M]. 北京：北京大学出版社,2006.
[10] 袁卫,庞皓,贾俊平,等. 统计学[M]. 4 版. 北京：高等教育出版社,2014.
[11] 于洪彦. Excel 统计分析与决策[M]. 北京：高等教育出版社,2001.
[12] 张勇,林军强. 财会统计业务基本技能训练教程[M]. 北京：中国人民大学出版社,2004.
[13] [美]布莱克,埃尔德雷奇. 以 Excel 为决策工具的商务与经济统计[M]. 张久琴,等,译. 北京：机械工业出版社,2003.
[14] 姜诗章. 统计学教程[M]. 北京：清华大学出版社,2006.
[15] 余华银. 统计学[M]. 长春：吉林大学出版社,2006.
[16] 曲岩,刘继运. 统计学[M]. 北京：北京大学出版社,中国林业出版社,2007.
[17] 刘馨. 统计学[M]. 成都：四川大学出版社,2006.
[18] 孙炎. 应用统计学[M]. 2 版. 北京：机械工业出版社,2014.
[19] 黎东升. 统计学原理[M]. 2 版. 北京：中国农业出版社,2013.
[20] 加拿大统计局《调查技能》项目组. 调查技能教程[M]. 北京：中国统计出版社,2002.
[21] 编写组. 社会经济统计学原理教科书[M]. 北京：中国统计出版社,1984.
[22] 柯惠新,丁立宏. 市场调查与分析[M]. 北京：中国统计出版社,2000.
[23] 刘思峰,吴和成,管利荣. 应用统计学[M]. 2 版. 北京：高等教育出版社,2011.
[24] 郭思亮,盛亦工,刘瑞娟,等. 统计学：方法与应用——以 Excel 为分析工具[M]. 成都：西南交通大学出版社,2014.
[25] 宋廷山,尉雪波,吴风庆. 应用统计学——以 Excel 为分析工具[M]. 成都：西南财经大学出版社,2006.
[26] 宋廷山,葛金田,王光玲. 统计学——以 Excel 为分析工具[M]. 2 版. 北京：北京大学出版社,2012.
[27] 宋廷山,郝克宁,张占涛. 经济计量模型：方法与应用——以统计应用软件为分析工具[M]. 成都：西南交通大学出版社,2012.
[28] 武松,潘发明. SPSS 统计分析大全[M]. 北京：清华大学出版社,2014.

教学支持说明

▶▶课件申请

尊敬的老师：

您好！感谢您选用清华大学出版社的教材！为更好地服务教学，我们为采用本书作为教材的老师提供教学辅助资源。鉴于部分资源仅提供给授课教师使用，请您直接手机扫描下方二维码实时申请教学资源。

任课教师扫描二维码
可获取教学辅助资源

▶▶样书申请

为方便教师选用教材，我们为您提供免费赠送样书服务。授课教师扫描下方二维码即可获取清华大学出版社教材电子书目。在线填写个人信息，经审核认证后即可获取所选教材。我们会第一时间为您寄送样书。

任课教师扫描二维码
可获取教材电子书目

清华大学出版社

E-mail: tupfuwu@163.com　　网址：http://www.tup.com.cn/

电话：8610-62770175-4506/4340　　传真：8610-62775511

地址：北京市海淀区双清路学研大厦B座509室　　邮编：100084